Matemáticas orientadas ás ensinanzas académicas

4ºB ESO
Capítulo 13: Estatística

www.apuntesmareaverde.org.es

Autora: Raquel Caro

Revisoras: María Molero e Nieves Zuasti

Tradutora: Mª Teresa Seara Domínguez

Revisora da tradución ao galego: Fernanda Ramos Rodríguez

Ilustracións: Banco de Imaxes de INTEF

Índice

A Estatística utilízase na Ciencia. Tamén para facer **sondaxes de opinión** como a aceptación polo público dun programa de televisión ou as enquisas sobre a intención de voto a un partido político. Úsanse técnicas estatísticas nos procesos de fabricación, é o **control de calidade**. Para facer previsións e programar o tráfico ou as necesidades de enerxía dun país. Cando se analiza un fenómeno observable aparecen unha serie de resultados que deben ser tratados convenientemente, de maneira que se poidan comprender mellor tanto os resultados como a característica obxecto de estudo correspondente a este fenómeno. Para este fin utilízase a Estatística.

Neste capítulo aprenderemos a recoñecer e clasificar distintos tipos de variables estatísticas, construír táboas de frecuencias e gráficos estatísticos para distintos tipos de variables estatísticas e determinar e interpretar medidas de centralización, posición e dispersión.

Tamén nos centraremos no estudo de dúas variables de interese correspondentes a dúas características (ou variables) distintas. Neste sentido, pode ser interesante considerar simultaneamente os dous caracteres a fin de estudar as posibles relacións entre eles.

1.FASES E TAREFAS DUN ESTUDO ESTATÍSTICO

Enfrontámonos a diario á necesidade de recoller, organizar e interpretar datos e esta necesidade aumentará no futuro, debido ao desenvolvemento dos sistemas de comunicación e das bases de datos. É notable o aumento do uso das redes sociais tales como *Youtube* ou *Facebook*, onde as persoas teñen oportunidade de presentar información sobre elas mesmas, e de páxinas web onde se poden encontrar

e descargar gran variedade de datos estatísticos sobre diversos temas da actualidade: resultados deportivos dos seus equipos favoritos, temperatura máxima e mínima ao longo dun mes, vendas de turrón o pasado Nadal, etc.

Noutras ocasións os datos son recollidos polo investigador mediante a realización dunha enquisa ou a través dun experimento. A enquisa requirirá a elaboración dun cuestionario, fixando os obxectivos do mesmo, elixindo as variables explicativas e redactando as preguntas que permitan obter a información desexada dunha forma clara e concisa.

Neste sentido, a estatística xogou un papel primordial neste desenvolvemento tecnolóxico que nos está tocando vivir, ao proporcionar ferramentas metodolóxicas xerais para analizar a variabilidade, determinar relacións entre variables, deseñar de forma óptima experimentos, mellorar as predicións e a toma de decisións en situacións de incerteza.

O tratamento estatístico dun problema comeza sempre coa presentación da magnitude que se quere analizar dunha determinada poboación e a selección da mostra pertinente para pasar á recollida de datos. Unha vez obtidos os datos ordénanse e preséntanse en táboas ou gráficas, de forma que sexa posible observar as particularidades que sinalan.

De aquí pódese considerar que un estudo estatístico consta dunha serie de fases e tarefas ben diferenciadas:

1. Definición da **poboación** e característica a estudar.

   Tarefas: Identificación das características cuantitativas e cualitativas; fixación da poboación; especificación da forma de recollida de datos (entrevistas, teléfono, correo electrónico, etc.).

2. Selección da **mostra**.

   Tarefas: Identificación do tamaño da mostra e orzamento necesario.

3. Recollida de **datos**.

   Tarefas: Deseño do cuestionario; deseño dunha mostra.

4. Organización e representación gráfica.

   Tarefas: Táboas e gráficas que axuden a unha máis fácil interpretación dos datos; isto consiste nun estudo de cada variable, a tabulación e representación(s) gráfica(s) máis apropiada(s).

5. Análise de datos.

   Tarefas: Tratamento dos datos. Isto consistirá nunha análise descritiva dos datos e/ou unha análise multivariante dos datos, dependendo do tipo de estudo a realizar e dos custes do mesmo.

6. Obtención de conclusións.

   Tarefas: recomendacións e toma de decisións a partir das conclusións.

Exemplo:

- Unha lista de puntos a ter en conta ao formular as preguntas da investigación é a seguinte:
  - Que queres probar? Que tes que medir /observar /preguntar?
  - Que datos necesitas? Como encontrarás os teus datos? Que farás con eles?
  - Cres que podes facelo? Encontrarás problemas? Cales?
  - Para que che servirán os resultados?

Desta maneira prepárase unha lista das características que queremos incluír no estudo, analizando as diferentes formas coas que poderían obterse os datos. Por simple observación: como o sexo, cor de pelo e ollos, se o alumno usa ou non lentes; se se require unha medición: como o peso, talle, perímetro de van; se habería que preguntar, é dicir, se se debe realizar unha enquisa: canto deporte practica, número do calzado, cantas horas dorme, cantas horas estuda ao día ou á semana, etc.

Polo tanto, é importante considerar a natureza das escalas de medida e tipo de variable estatística, xa que delas depende o método de análise de datos que se pode aplicar. A elección do conxunto de datos é crítica pois, dependendo do tipo de datos, a gama de técnicas estatísticas será máis ou menos ampla, xa que non todas as técnicas son aplicables a calquera tipo de dato.

2. POBOACIÓN E MOSTRA. VARIABLES ESTATÍSTICAS

2.1. Poboación

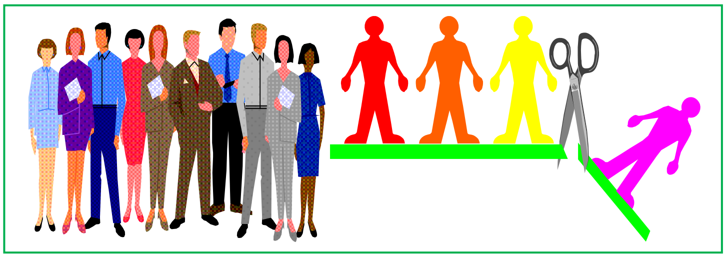
Poboación estatística, colectivo ou universo é o conxunto de todos os individuos (persoas, obxectos, animais, etc.) que conteñan información sobre o fenómeno que se estuda.

Exemplos:

- Se estudamos o prezo da vivenda nunha cidade, a poboación será o total das vivendas desta cidade.
- Vaise realizar un estudo estatístico sobre a porcentaxe de persoas casadas na península. Para iso non é factible estudar todos e cada un dos habitantes por razóns de custe e de rapidez na obtención da información. Polo tanto é necesario examinar só unha parte desta **poboación**. Esa parte é a **mostra** elixida.

1.2. Mostra

Mostra é un subconxunto representativo que se selecciona da poboación e sobre o que se vai realizar a análise estatística. O **tamaño da mostra** é o número dos seus elementos. Cando a mostra comprende todos os elementos da poboación denomínase **censo.**



Exemplo:

- Se se estuda o prezo da vivenda dunha cidade, o normal será non recoller información sobre todas as vivendas da cidade (xa que sería un labor moi complexo e custoso) senón que se soe seleccionar un subgrupo (mostra) que se entenda que é suficientemente representativo.

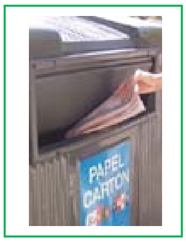
Actividades propostas

1. Sinalar en que caso é máis conveniente estudar a poboación ou unha mostra:

a) O diámetro dos parafusos que fabrica unha máquina diariamente.

b) A altura dun grupo de seis amigos.

2. Pódese ler o seguinte titular no xornal que publica o teu instituto: "*A nota media dos alumnos de 4º ESO da Comunidade de Madrid é de* 7.9". Como se chegou a esta conclusión? Estudouse toda a poboación? Se tivesen seleccionado para o seu cálculo só ás mulleres, sería representativo o seu valor?



2.3. Individuo ou unidade estatística

Individuo ou **unidade estatística** é calquera elemento que conteña información sobre o fenómeno que se estuda.

Exemplo:

- Se estudamos as notas dos alumnos dunha clase, cada alumno é un individuo; se estudamos o prezo da vivenda, cada vivenda é unha unidade estatística.

2.4. Variable estatística

En xeral, suporemos que se está analizando unha determinada poboación, da que nos interesa certa característica, representada por unha **variable** observable ou estatística *X*. As variables que están baixo estudo pódense clasificar en dúas categorías:

Variables cualitativas ou atributos (datos non métricos), que non se poden medir numericamente. As escalas de medida non métricas clasifícanse en nominais (o categóricas) e ordinais.

Variables cuantitativas, que teñen un valor numérico. Este tipo de variables son as que aparecen con máis frecuencia e permiten unha análise máis detallada que as cualitativas. Dentro das variables cuantitativas, pódense distinguir as variables **discretas** e as variables **continuas**. As **variables discretas** toman valores illados, mentres que as **variables continuas** poden tomar calquera valor dentro dun intervalo.

Exemplo:

- Exemplos de variables cualitativas son a nacionalidade ou a raza dun conxunto de persoas.
- Exemplos de variables cuantitativas son as notas obtidas nunha materia, o peso ou a altura dun conxunto de persoas.
- Exemplos de variables discretas son o número de alumnos que aproban unha materia ou o número de compoñentes defectuosos que se producen ao día nunha fábrica.
- Exemplos de variables continuas son o tempo que tardamos en chegar ao instituto desde a nosa casa ou a velocidade dun vehículo.

Actividades resoltas

- Vaise realizar un estudo estatístico sobre a porcentaxe de persoas con fillos nunha localidade madrileña de 134 678 habitantes. Para iso elíxense 2 346 habitantes e esténdense as conclusións a toda a poboación. Identificar: variable estatística, poboación, mostra, tamaño dunha mostra e individuo.
    - *Variable estatística*: se unha persoa ten fillos ou non.
    - *Poboación*: os 134 678 habitantes da localidade.
    - *Mostra*: os 2 346 habitantes elixidos.
    - *Tamaño dunha mostra:* 2 346 persoas.
    - *Individuo*: Cada persoa á que se lle pregunte.

Actividades propostas

3. Indica o tipo de variable estatística que estudamos e razoa, en cada caso, se sería mellor analizar unha mostra ou a poboación:
   a) O sexo dos habitantes dun país.
   b) O diñeiro gastado á semana polo teu irmán.
   c) A cor de pelo dos teus compañeiros de clase.
   d) A temperatura da túa provincia.
   e) O talle de pé dos alumnos do instituto.

4. Para realizar un estudo facemos unha enquisa entre os mozos dun barrio e preguntámoslles polo número de veces que van ao cine ao mes. Indica que características debería ter a mostra elixida e se deberían ser todos os mozos da mostra da mesma idade.

3. TÁBOAS DE FRECUENCIAS

3.1. Frecuencia absoluta

Cando se analiza unha *variable discreta*, a información resultante da mostra encóntrase resumida habitualmente nunha táboa ou distribución de frecuencias. Supoñamos que se tomou unha mostra de tamaño N na que se identificaron k valores (ou modalidades) distintos $x_1, x_2, \ldots, x_k$. Cada un deles prodúcese cunha **frecuencia absoluta** n_i, é dicir, o número de veces que aparece na mostra.

A información obtida pódese resumir nunha **táboa de frecuencias**.

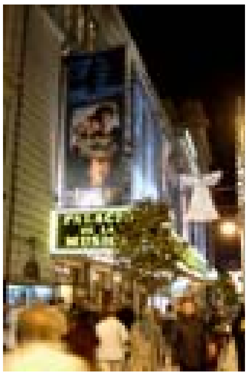
As táboas de frecuencia tamén se utilizan para representar información dunha *variable continua* procedente dunha mostra na que se agrupan as observacións en intervalos que se denominan **intervalos de clase** L_i ou celas.

Aínda que este procedemento supón, de feito, unha perda de información, esta perda non é de magnitude importante e vese compensada coa agrupación da información e a facilidade de interpretación que proporciona unha táboa de frecuencias.

Neste caso, os valores x_i correspóndense co punto medio do intervalo e denomínanse **marcas de clase**.

Exemplo:

- Cando realizamos un estudo sobre o ocio e enquisamos a 40 mozos dunha localidade sobre o número de veces que van ao cine os resultados desta enquisa podémolos recoller nunha táboa para resumir esta información.



Actividades resoltas

- Estase realizando un control do peso dun grupo de nenos. Para iso, contabilízase o número de veces que comen ao día unha barra de chocolate 13 nenos durante un mes, obtendo os seguintes números: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2.

A información obtida pódese resumir nunha táboa de frecuencias absolutas:

Valores	0	1	2	3	4	5	6	7
Frecuencia absoluta	2	2	4	1	2	1	0	1

- Nunha fábrica realízase un estudo sobre o espesor, en *mm*, dun certo tipo de latas de refresco. Con este fin, selecciona unha mostra de tamaño N = 25, obtendo os seguintes valores: 7.8, 8.2, 7.6, 10.5, 7.4, 8.3, 9.2, 11.3, 7.1, 8.5, 10.2, 9.3, 9.9, 8.7, 8.6, 7.2, 9.9, 8.6, 10.9, 7.9, 11.1, 8.8, 9.2, 8.1, 10.5.

Esta información pódese resumir na seguinte táboa de frecuencias, con 5 intervalos: (7, 8], (8, 9], (9, 10], (10, 11], (11, 12], sendo as marcas de clase os puntos medios de cada intervalo: 7.5; 8.5; 9.5; 10.5; 11.5. Comproba que as frecuencias absolutas son as indicadas na táboa:

Intervalos de clase	**(7, 8]**	**(8, 9]**	**(9, 10]**	**(10, 11]**	**(11, 12]**
Marcas de clase	7.5	8.5	9.5	10.5	11.5
Frecuencia absoluta	6	8	5	4	2

Actividades propostas

5. Obter a táboa de frecuencias absolutas das notas en inglés de 24 alumnos:

6	6	7	8	4	9	8	7	6	5	3	5
7	6	6	6	5	4	3	9	8	8	4	5

3.2. Frecuencia relativa

Denomínase **frecuencia relativa** (f_i) dun valor da variable ao cociente entre a frecuencia absoluta e o número total de observacións N. Escríbese:

$$f_i = \frac{n_i}{N} \leq 1$$

Exemplo:

- Da mesma maneira podemos recoller a información obtida a partir dunha enquisa a 40 mozos dunha localidade sobre o número de veces que van ao cine mediante porcentaxe do número de veces que se repite un valor da variable sobre o total.

Actividades resoltas

- Estase realizando un control do peso dun grupo de nenos. Para iso, contabilízase o número de veces que comen ao día unha barra de chocolate 13 nenos durante un mes, obtendo os seguintes números: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2.

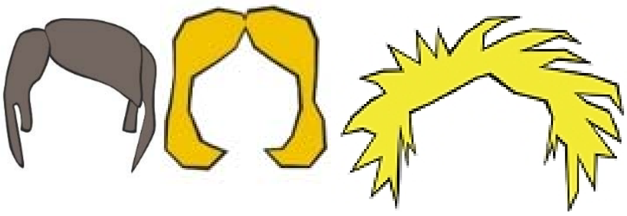
A información obtida pódese resumir nunha táboa de frecuencias relativas:

Valores	0	1	2	3	4	5	6	7
Frecuencia relativa	0.154	0.154	0.307	0.077	0.154	0.077	0	0.077

Actividades propostas

6. Construír unha táboa de frecuencias relativas coa cor de pelo de 24 persoas elixidas ao azar:
M =moreno; L = louro; P = pelirroxo

M	L	P	L	L	L
L	P	P	M	M	M
M	R	L	L	L	L
M	M	M	M	M	P



3.3. Frecuencia absoluta acumulada

Denomínase **frecuencia absoluta acumulada** dun valor da variable N_i á suma de todas as frecuencias absolutas dos valores menores ou iguais ca el. Calcúlase como:

$$N_i = \sum_{j=1}^{i} \left[n_j \right]$$

Verifícase a seguinte relación entre os valores de N_i:

$$N_1 \leq N_2 \leq \cdots \leq N_k = N$$

Exemplo:

- Da mesma maneira podemos recoller a información obtida a partir dunha enquisa a 40 mozos dunha localidade sobre o número de veces que van ao cine mediante o número acumulado de veces que se repite un valor da variable sobre o total.

Actividades resoltas

- Estase realizando un control do peso dun grupo de nenos. Para iso, contabilízase o número de veces que comen ao día unha barra de chocolate 13 nenos durante un mes, obtendo os seguintes números: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2.

A información obtida pódese resumir nunha táboa de frecuencias absolutas:

Valores	0	1	2	3	4	5	6	7
Frecuencia absoluta	2	2	4	1	2	1	0	1
Frecuencia absoluta acumulada	2	4	8	9	11	12	12	13

Actividades propostas

7. O número de horas diarias de estudo de 14 alumnos é o seguinte:

3 4 2 5 3 4 3 2 3 4 5 4 3 2

a) Efectúa un reconto e organiza os resultados obtidos nunha táboa de frecuencias absolutas acumuladas.

b) Que significan as frecuencias acumuladas que calculaches?

3.4. Frecuencia relativa acumulada

Denomínase **frecuencia relativa acumulada** (F_i) dun valor da variable á suma de todas as frecuencias relativas dos valores menores ou iguais ca el. Calcúlase como:

$$F_i = \sum_{j=1}^{i} \left[f_j \right]$$

Verifícase a seguinte relación entre os valores de F_i:

$$F_1 \le F_2 \le \cdots \le F_k = 1$$

Exemplo:

- Da mesma maneira podemos recoller a información obtida a partir dunha enquisa a 40 mozos dunha localidade sobre o número de veces que van ao cine mediante a porcentaxe acumulada do número de veces que se repite un valor da variable sobre o total.

Actividades resoltas

- Estase realizando un control do peso dun grupo de nenos. Para iso, contabilízase o número de veces que comen ao día unha barra de chocolate 13 nenos durante un mes, obtendo os seguintes números: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2.

A información obtida pódese resumir nunha táboa de frecuencias relativas:

Valores	0	1	2	3	4	5	6	7
Frecuencia relativa	0.154	0.154	0.307	0.077	0.154	0.077	0	0.077
Frecuencia relativa acumulada	0.154	0.308	0.615	0.692	0.846	0.923	0.923	1

- Nunha fábrica realízase un estudo sobre o espesor, en *mm*, dun certo tipo de latas de refresco. Con este fin, selecciona unha mostra de tamaño $N = 25$, obtendo os seguintes valores: 7.8, 8.2, 7.6, 10.5, 7.4, 8.3, 9.2, 11.3, 7.1, 8.5, 10.2, 9.3, 9.9, 8.7, 8.6, 7.2, 9.9, 8.6, 10.9, 7.9, 11.1, 8.8, 9.2, 8.1, 10.5.

Esta información pódese resumir na seguinte táboa de frecuencias, con 5 intervalos:

Intervalos de clase	**(7, 8]**	**(8, 9]**	**(9, 10]**	**(10, 11]**	**(11, 12]**
Marcas de clase	7.5	8.5	9.5	10.5	11.5
Frecuencia absoluta	6	8	5	4	2
Frecuencia relativa	0.24	0.32	0.2	0.16	0.08
Frecuencia relativa acumulada	0.24	0.56	0.76	0.92	1

- Organízase nunha táboa a información recollida das estaturas, en cm, dun grupo de 20 nenas:

130	127	141	139	138	126	135	138	134	131
143	140	129	128	137	136	142	138	144	136

A estatura é unha variable estatística cuantitativa continua. Polo tanto, podemos agrupar os valores da variable en intervalos que chamamos clases ou celas. A amplitude de cada intervalo vén dada pola fórmula:

$$\frac{Máx - Mín}{\sqrt{N}}$$

No noso caso concreto temos que:

$$\frac{144-126}{\sqrt{20}} = 4.02$$

Aproximando, a amplitude de cada intervalo é de 5 cm.

Estatura en intervalos	[125-130)	[130-135)	[135-140)	[140-145)
Frecuencia absoluta	4	3	8	5
Frecuencia relativa	0.2	0.15	0.4	0.25
Frecuencia absoluta acumulada	4	7	15	20
Frecuencia relativa acumulada	0.2	0.35	0.75	1

Actividades propostas

8. Nunha avaliación, dos 30 alumnos dunha clase, o 30 % aprobou todo, o 10 % suspendeu unha materia, o 40 % suspendeu dúas materias e o resto máis de dúas materias.

a) Realiza a táboa de frecuencias completa correspondente (frecuencias absolutas, frecuencias relativas, frecuencias absolutas acumuladas e frecuencias relativas acumuladas).

b) Hai algún tipo de frecuencia que corresponda á pregunta de cantos alumnos suspenderon menos de dúas materias? Razoa a resposta.

4. GRÁFICOS ESTATÍSTICOS

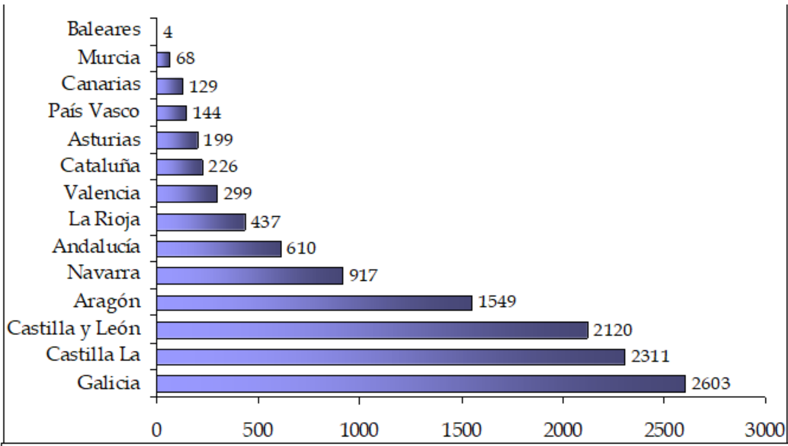
4.1. Diagrama de barras

Existen numerosas maneiras de representar graficamente a información que se obtivo dunha mostra, dependendo do tipo de variable que se estea analizando e do fin que se persiga coa representación.

Cando se quere representar graficamente unha variable cualitativa (atributo) ou unha variable cuantitativa discreta pódense utilizar os ***diagramas de barras ou rectángulos***. Colócanse os valores da variable (as modalidades do atributo ou valores da variable discreta) no eixe de abscisas e, no eixe de ordenadas, as frecuencias (absolutas ou relativas). Sobre cada valor levántase unha barra ou rectángulo cuxa altura é igual á frecuencia. Por comodidade, ás veces tamén se soen intercambiar os eixes.

Exemplo:

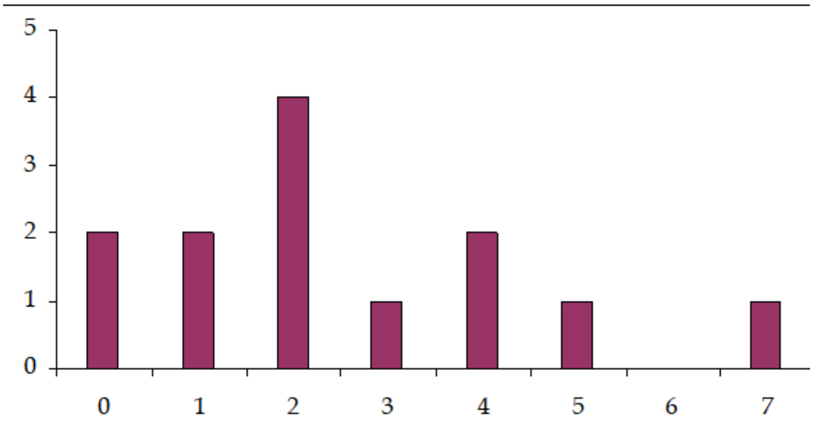
- Representouse graficamente a potencia eólica (fonte de enerxía eléctrica renovable) instalada en España por Comunidade Autónoma en xaneiro de 2014 (en Megavatios)



DATOS 2014 da AEE

Exemplo:

- Representouse graficamente o número de fallos mensuais dunha máquina de xeados



Número de fallos mensuais

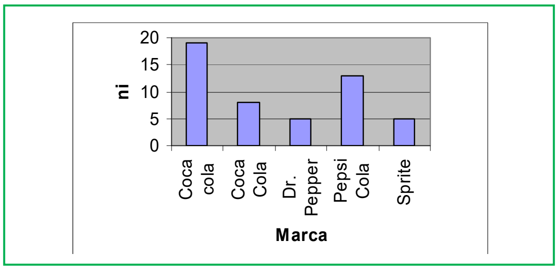
Actividades resoltas

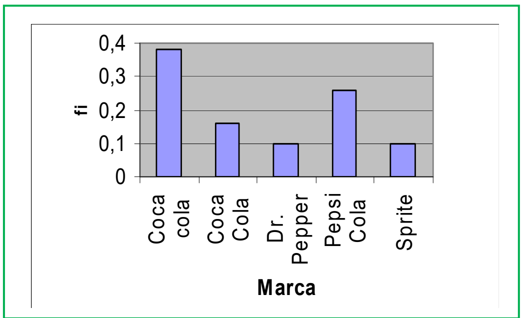
- Dada a seguinte información correspondente ás preferencias de 50 adolescentes americanos respecto á marca de refresco que consomen, construe a táboa asociada a estes datos e represéntaos graficamente nun diagrama de barras de frecuencias absolutas e noutro de frecuencias relativas.

COCA-COLA = CC; COCA-COLA LIGHT = CCL; DR.PEPPER = A; PEPSI-COLA = PC, SPRITE = S

CCL	CC	S	A	CC	CC	A	CC	P	CC
S	CCL	P	CCL	CC	CC	CCL	P	P	A
S	S	CC	CC	CC	A	P	CC	CCL	CC
CCL	CC	P	P	P	CCL	P	S	P	CC
CC	P	CCL	CC	CC	P	CC	P	CC	A

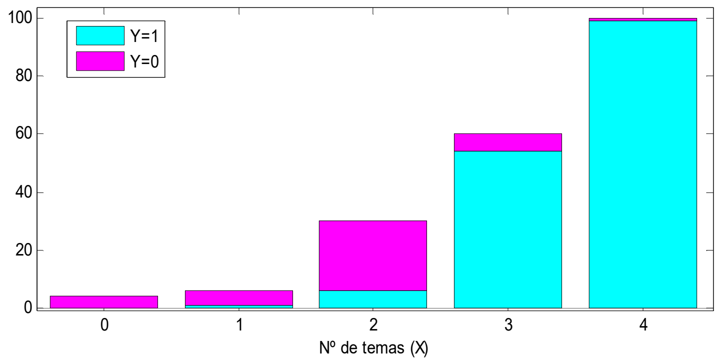
Marca	n_i	f_i
Coca Cola	19	0,38
Coca Cola Light	8	0,16
Dr. Pepper	5	0,10
Pepsi Cola	13	0,26
Sprite	5	0,10
	50	1

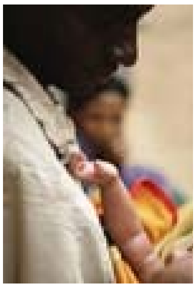




Actividades propostas

9. Se queremos representar conxuntamente valores da variable correspondentes a diferentes períodos de tempo, ou a distintas calidades, para comparar situacións podemos construír un diagrama de barras apiladas. Poderías interpretar este gráfico correspondente ao número de temas que os alumnos dunha materia de 4º ESO levan estudados? Tómase información en dúas clases dun instituto (azul e rosa).





10. O sexo de 18 bebés nacidos nun hospital de Madrid foi:

H	M	H	H	M	H
H	M	M	H	M	H
M	M	H	H	M	H

Constrúe a táboa asociada a estes datos e represéntaos.

11. Representa os valores da variable da táboa adxunta co gráfico adecuado correspondentes a unha enquisa realizada sobre o sector ao que pertence nun grupo de traballadores madrileños.

SECTOR	INDUSTRIAL	AGRARIO	SERVIZOS	OUTROS
% TRABALLADORES	20	16	45	19

4.2. Histogramas

A representación máis utilizada en variables cuantitativas continuas é o **histograma**.

No eixe de abscisas colócanse os diferentes intervalos nos que se agrupan as observacións da variable. Sobre estes intervalos, levántanse rectángulos cuxa **área** é proporcional á frecuencia observada en cada un deles.

No caso de que todos os intervalos teñan a mesma amplitude basta con que a altura dos rectángulos sexa proporcional á frecuencia.

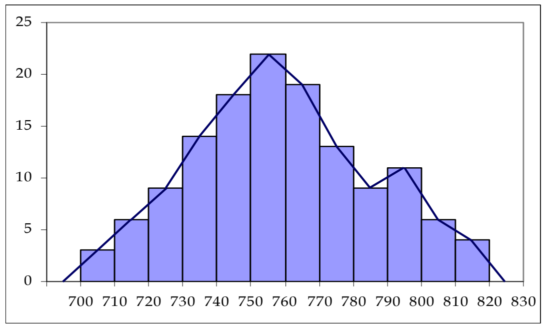
Dependendo das frecuencias que se utilicen, tratarase dun histograma de frecuencias relativas ou ben dun histograma de frecuencias absolutas.

En ocasións, únense os puntos medios dos segmentos superiores dos rectángulos, obténdose deste modo o **polígono de frecuencias**, xa sexan absolutas ou relativas. Estes polígonos constrúense utilizando un intervalo anterior ao primeiro (da mesma lonxitude ca este) e outro posterior ao último (da súa mesma lonxitude). Desta maneira, os polígonos delimitan unha área pechada.

En ambos os casos, tamén se poden utilizar as frecuencias acumuladas para construír os respectivos histogramas. Estes histogramas tamén levan asociados os correspondentes polígonos de frecuencias, que, neste caso, se constrúen unindo os vértices superiores dereitos de cada un dos intervalos.

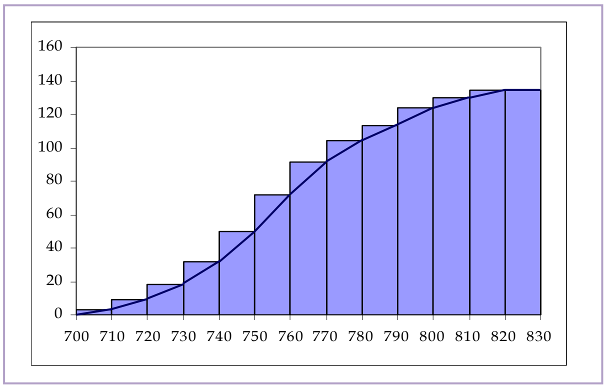
Exemplo:

- Representouse graficamente a información obtida a partir das emisións específicas de CO_2 dunha central de carbón (kg/megavatio-hora) a partir dun histograma e dun polígono de frecuencias absolutas.



Exemplo:

- Representouse graficamente a información obtida a partir das emisións específicas de CO_2 dunha central de carbón (kg/megavatio-hora) a partir dun histograma e dun polígono de frecuencias acumuladas absolutas.



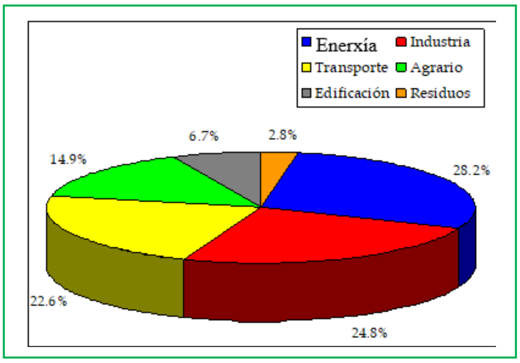
Actividades propostas

12. Completa a táboa de frecuencias para poder representar a información mediante o histograma de frecuencias acumuladas:

IDADE	[15, 25)	[25, 35)	[35, 45)	[45, 55)
NÚMERO DE PERSOAS	25	45	55	65

13. A que representación gráfica corresponde o seguinte gráfico correspondente á información recollida sobre a idade de 100 persoas? Por que cres que se utilizou este e non outro?

4.3. Diagrama de sectores



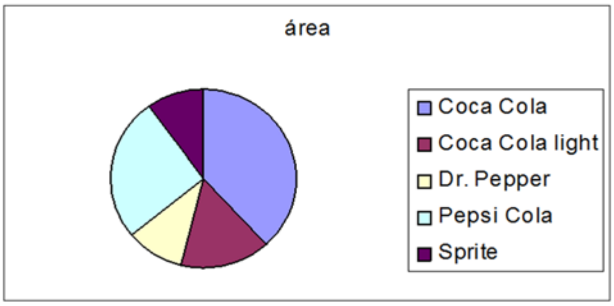
No **diagrama de sectores** colócanse as modalidades do atributo (variable cualitativa) ou valores dunha variable cuantitativa discreta nun círculo, asignando a cada un un sector do círculo de **ángulo proporcional á súa frecuencia**. Non resulta moi operativo cando a variable ten demasiadas categorías.

Exemplo:

Da mesma maneira podemos recoller a información obtida de emisións de gases de efecto invernadoiro en España no período 1999-2012 (%)

Actividades resoltas

- Dada a información correspondente ás preferencias de 50 adolescentes americanos respecto á marca de refresco que consomen da actividade resolta do apartado 3.1. realizar o gráfico de sectores.



Actividades propostas

14. Dos 100 asistentes a unha voda, o 34 % comeu tenreira de segundo prato, 25 % pato, 24 % año e o resto peixe.

a) Organiza a información anterior nunha táboa de frecuencias e representa os datos nun gráfico de sectores.

b) Realiza un diagrama de barras e explica como o fas. Cal dos dous gráficos prefires? Por que?

15. Recolleuse información sobre o contido de sales minerais de 24 botellas de auga dun grupo de escolares nunha excursión tal que:

45	45	65	56	33	65	23	23
34	23	43	67	22	43	34	23
12	34	45	34	19	34	23	43

a) Clasifica a variable estatística estudada.

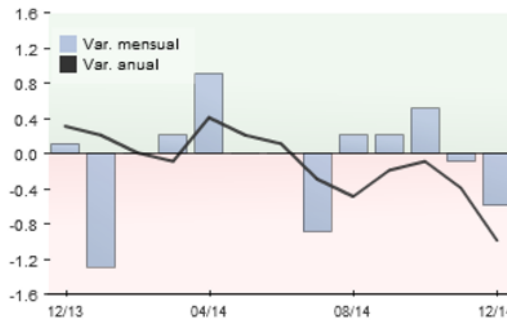
b) Sería conveniente tomar ou non intervalos ao facer unha táboa de frecuencias?

c) Realiza o gráfico que consideres máis oportuno.

4.3. Análise crítica de táboas e gráficas estatísticas nos medios de comunicación. Detección de falacias

Os medios de comunicación recorren con frecuencia a táboas e gráficas que axuden a unha máis doada interpretación dos datos por parte do público en xeral. Un caso pode ser o seguinte gráfico que presenta o Instituto Nacional de Estatística (INE), que representa o índice dos prezos ao consumo.

Índice de prezos de consumo. IPC.



Índice xeral nacional (%)

Porén, non é raro observar como se utilizan uns mesmos datos estatísticos para obter conclusións distintas.

- Unha suba de prezos ou do índice de desemprego pode parecer máis ou menos acentuada segundo quen presente a información.
- Un índice de audiencia ou o colesterol dun determinado alimento poden parecer máis ou menos altos segundo con que sexa comparado.
- As chamadas telefónicas parecen ser máis baratas nunha compañía que noutra.

A lista de exemplos é interminable.

Deste modo, a Estatística, ademais do papel instrumental que presentamos ata agora, ten un importante papel no desenvolvemento do pensamento crítico que nos manterá atentos a estes excesos.

Os erros máis frecuentes, aínda que ás veces non se trata de erros senón de manipulacións tendenciosas, son os seguintes:

- Erros na obtención de datos.
- Limitacións humanas ou dos instrumentos: é imposible, por exemplo, medir o peso ou a estatura dunha persoa con infinita precisión. Pero mesmo en estudos exhaustivos, como os censos, estímanse os erros de mostraxe.
- Cuestionarios mal formulados: se non se recollen todas as posibles respostas, se a pregunta inflúe na resposta, se as preguntas conteñen xuízos de valor ou se as diferentes opcións de resposta non son equilibradas (por exemplo: si, ás veces, non). O conxunto de respostas posibles pode facer que haxa duplicacións u omisións. Incorrer neste erro, deliberadamente ou non, deixa a individuos da poboación sen representación entre as respostas e, polo tanto, os resultados que saian do estudo estarán nesgados. As modalidades da variable deben ser incompatibles e exhaustivas (por exemplo: se preguntamos pola cor favorita e ofrecemos como posibles respostas "Vermello", "Azul" ou "Amarelo", deixamos sen poder responder aos que queren escoller outra cor; se non estamos interesados noutras cores, podemos incluír un apartado chamado "Outra").
- Delimitación imprecisa da poboación: Por exemplo, se se desexa estudar se os nenos madrileños ven demasiado a televisión, haberá que deixar claro que idades en concreto se considerarán, se entendemos por madrileño a calquera residente ou só aos nacidos en Madrid, etc.
- Selección da mostra non apropiada ou non representativa: a mostra non representa á poboación. A elección dos individuos concretos que forman parte da mostra debe facerse de forma aleatoria. Por exemplo: se estudamos os gustos televisivos dos adolescentes dun instituto e pensamos que estes gustos poden variar en función da idade, na selección da mostra deben escollerse idades variadas, a poder ser, na mesma proporción na que se presentan no instituto.
- Erros nas táboas: os datos non están ordenados, evitar ambigüidades nos extremos dos intervalos para variables continuas, etc.
- Erros nas gráficas: nos diagramas de barras falta a orixe, están truncados ou hai erros na escala dos eixes, etc. Hai que deixar claras as variables que se miden.
- Erros nos parámetros de medida: por exemplo a media non é representativa (poboacións heteroxéneas) ou vese afectada por valores moi grandes; confusión entre media e mediana.
- Erros nos pictogramas con superficies onde se inscriben proporcionais ao cadrado das frecuencias.

5. MEDIDAS DE TENDENCIA CENTRAL

As medidas de tendencia central ou de centralización son as que, intuitivamente, aparecen en primeiro lugar ao intentar describir unha poboación ou mostra.

Pódense dividir en tres clases: **medidas de tamaño, de frecuencia e de posición**.

No que segue, suporemos que estamos analizando unha poboación da que se toma unha mostra de tamaño *N*, é dicir, que está composta por *N* individuos (ou observacións), dos cales se desexa estudar a variable *X*, o que dá lugar á obtención de *N* valores que se representan por $x_1, x_2, \ldots, x_N$. Estes valores non se supoñen ordenados senón que o subíndice indica a orde na que foron seleccionados.

5.1. Medidas de tamaño

As medidas de tamaño defínense a partir dos valores da mostra, así como da súa frecuencia.

Definimos así a **media aritmética** ou **termo medio** ou, simplemente, **media** como:

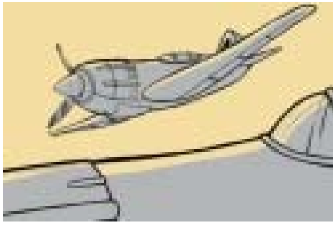
$$\overline{x} = \frac{\sum_{i=1}^{N}[x_i]}{N}$$

Pódese interpretar como o centro de masas das observacións da mostra. Dentro das súas vantaxes pódese destacar que utiliza todas as observacións, que son facilmente calculables, teñen unha interpretación sinxela e boas propiedades matemáticas. O seu inconveniente é que se pode ver afectada polos valores anormalmente pequenos ou grandes que existan na poboación ou mostra (denominados *outliers*).

No caso de que teñamos unha variable cuantitativa agrupada en intervalos o valor da variable *X* que representa ao intervalo para poder calcular a media aritmética é a **marca de clase** e calcúlase como a semisuma dos valores extremos do intervalo.

Exemplo:

- Recóllese a información referida ao número de horas de voo diarias de 20 azafatas. Se a media é igual a 4.1, isto indica que, por termo medio, o número de horas de voo é 4.1.



Exemplo:

- Da mesma maneira se recollemos a información sobre a idade media da túa clase obteremos un valor entre 15 e 16 anos. A idade media será por exemplo 15.4, valor teórico, que pode non coincidir con ningún dos valores reais.

Actividades resoltas

- Un fabricante de xeados está realizando un control de calidade sobre certas máquinas respecto á súa capacidade de regular a temperatura de refrixeración. Para iso, selecciona unha mostra de *N*=16 máquinas da fábrica e mide con precisión o valor da súa capacidade (na unidade de medida μF), obtendo os seguintes resultados: 20.5, 19.8, 19.6, 19.2, 23.5, 28.9, 19.9, 19.2, 20.1, 18.8, 19.5, 20.2, 18.6, 19.7, 22.1, 19.3. Utilizando estes valores de capacidade, obter a media aritmética.

$$\overline{x} = \frac{\sum_{i=1}^{N}[x_i]}{N} = \frac{20.5+19.8+19.6+19.2+23.5+28.9+19.9+19.2+20.1+18.8+19.5+20.2+18.6+19.7+22.1+19.3}{16} = \mathbf{20.56} \quad \mu\mathbf{F}$$

Actividades propostas

16. Unha persoa ingresa 10 000 euros nun fondo de inversión o 1 de xaneiro de 2009. As rendibilidades anuais do fondo durante os anos seguintes foron as seguintes:

Ano	2009	2010	2011	2012
Rendibilidades (%)	5	3	–1	4

Se non retirou o capital, cal foi a rendibilidade media do fondo durante estes anos?

17. Interpreta os valores da variable desta táboa que representa o peso de 100.000 bombonas de butano dunha fábrica, en quilogramos. Que gráfico utilizarías? Calcula a media e interprétaa.

Peso [ )	**f_i %**	**n_i**	**N_i**
14.5 - 15	0.3	300	300
15 - 15.5	1.6	1 600	1 900
15.5 - 16	7.4	7 400	9 300
16 - 16.5	21.5	21 500	30 800
16.5 - 17	30.5	30 500	61 300
17 - 17.5	24.5	24 500	85 800
17.5 - 18	10.7	10 700	96 500
18 - 18.5	21.5	21 500	30 800

5.2. Medidas de frecuencia

Defínense tendo en conta unicamente a frecuencia dos valores da variable da mostra.

A **moda** (*Mo*) defínese como o valor da variable que se obtivo con maior frecuencia. Pode haber máis dunha moda.

Exemplo:

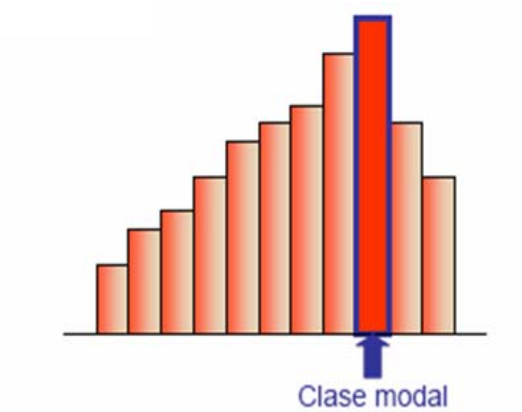
- Realízase un estudo entre 200 espectadores a un musical en Madrid para determinar o grao de satisfacción, obténdose os seguintes resultados:

Opinión	Moi bo	Bo	Regular	Malo	Moi malo
%	75	25	45	15	40

A modalidade que máis se repite é "moi bo", polo que a moda é Mo = Moi bo.

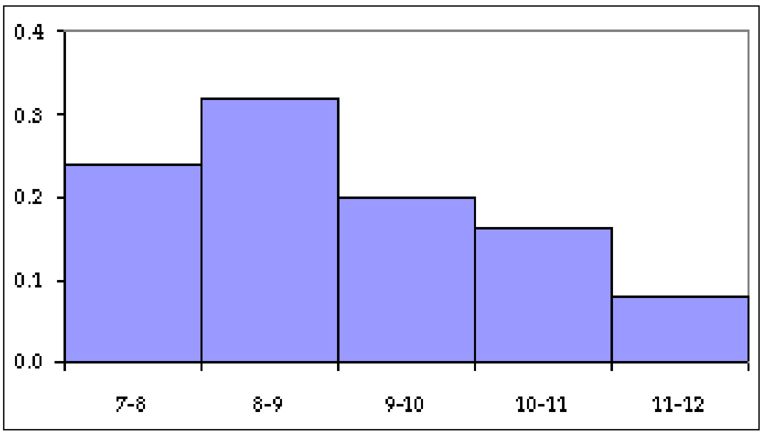
Exemplo:

- No caso de que a distribución estea agrupada en intervalos haberá que identificar a clase modal, é dicir, o intervalo onde hai maior número de valores da variable.



Actividades resoltas

- A partir da táboa de frecuencias do espesor de latas de refresco, podemos debuxar os seus histogramas de frecuencias relativas e determinar onde está a súa moda. É dicir no intervalo [8-9). A moda sinala que o máis frecuente é ter un espesor entre 8 e 9 mm.



Actividades propostas

18. Obter a media e a moda dos seguintes valores da variable referidos ao resultado de lanzar un dado 50 veces.

1	2	3	2	3	4	3	3	3	5
5	5	5	6	5	6	5	6	4	4
3	2	1	2	3	4	5	6	5	4
3	2	3	4	5	6	5	4	3	2
3	4	5	5	5	5	6	6	6	3

19. Realizar a actividade anterior pero agrupando en intervalos de amplitude 2, empezando en 0. Obtés os mesmos resultados? Por que?

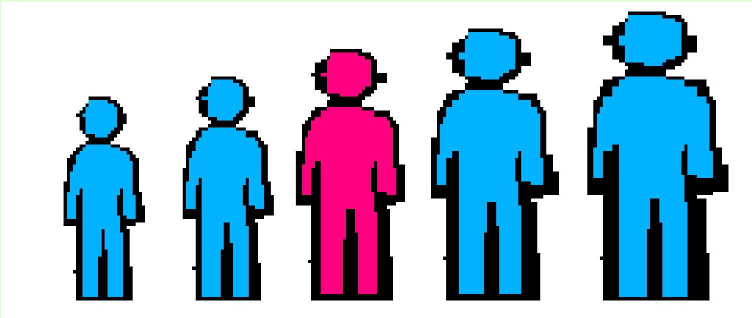
5.3. Medidas de posición

Defínense a partir da posición dos valores da mostra.

En xeral, coñécense co nome de **centís** ou **percentís**.

Se reordenamos en orde crecente os valores tomados da mostra e os denotamos por $x_{\{1\}}$, $x_{\{2\}}$,..., $x_{\{N\}}$ pódense definir as seguintes medidas de posición:

- ✓ A ***mediana*** *Me* é un valor tal que o 50 % das observacións son inferiores a el. Non ten por que ser único e pode ser un valor non observado.



Altura mediana

- ✓ Os ***cuartís*** (ou cuartilas) Q_1, Q_2 e Q_3 son os valores tales que o 25 %, 50 % e 75 % (respectivamente) dos valores da variable son inferiores a el.
- ✓ Os ***decís*** D_1, D_2,..., D_9 son os valores tales que o 10 %, 20 %, … , 90 % (respectivamente) dos valores da variable son inferiores a el.

En xeral, defínese o **percentil** ou **centil** do *k* % (sendo $0 \leq k \leq 100$) como o valor tal que o *k* % das observacións son inferiores a el.

A mediana e o resto de medidas de posición teñen como principal vantaxe a súa fácil interpretación e a súa robustez (non se ven afectadas por observacións extremas).

Exemplo:

- Calcula os cuartís e o percentil 65 dos seguintes valores da variable referidos ao número de fillos das familias dun bloque de edificios da localidade de Madrid:

Número de fillos	f_i	F_i
1	11	11
2	27	38
3	4	42
4	18	60
Total	60	

Para calcular o primeiro cuartil calculamos o 25 % do total dunha mostra $N = 60$, é dicir, $60 \cdot 0.25 = 15$. Así, o primeiro cuartil ten 15 valores da variable menores e o resto maiores. Na columna de frecuencias acumuladas, o primeiro número maior ou igual que 15 é 38, que corresponde ao valor da variable 2. Polo tanto o primeiro cuartil é 2 (ou con mellor aproximación un valor entre 1 e 2). Da mesma forma o 50 % de 60 é 30, é dicir o cuartil 2 (Mediana) sería tamén 2 (ou de novo, un valor entre 1 e 2). O 75 % de 60 sería 45 e desta forma o cuartil 3 sería 4 (ou un valor entre 3 e 4) posto que o valor maior a 45 é 60, que corresponde ao valor 4 da variable obxecto de estudo. Por último, o percentil 65 corresponde ao valor 3 xa que 65 % de 60 é igual a 39 e o valor maior que 39 é 42.

Resumo:

25 % de 60 = 15 → 38 > 15 > 11 → $Q_1 = 2$

50 % de 60 = 30 → 38 > 30 > 11 → Me = $Q_2 = 2$

75 % de 60 = 45 → 60 > 45 > 42 → $Q_3 = 4$

65 % de 60 = 39 → 42 > 39 > 38 → $P_{65} = 3$

As medidas de posición permítennos realizar outro tipo de gráfico estatístico que se chama o ***gráfico de caixa***.

Para realizar este gráfico, constrúese unha *caixa* (xa sexa horizontal ou vertical), cuxos lados coinciden co *primeiro e terceiro cuartil* Q_1 e Q_3. Polo tanto, a caixa abrangue o 50% das observacións realizadas. Dentro desta caixa, inclúese un segmento (ou ben un punto) que corresponde á *mediana*.

De cada lado da caixa parte un segmento que se estende ata os valores correspondentes ás observacións *mínima e máxima* $x_{\{1\}}$ e $x_{\{N\}}$.

Actividades resoltas

- Estase realizando un control de calidade sobre os fallos dunhas determinadas máquinas. Para iso, contabilízanse os fallos de N=13 máquinas durante un mes, obtendo os seguintes números de fallos: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2. Utilizando estes valores obter as medidas de tendencia central e resumir nunha táboa de frecuencias a información obtida do número de fallos mensuais das máquinas, obtendo a media aritmética doutra maneira.

$$\bar{x} = \frac{\sum_{i=1}^{N}[x_i]}{N} = \frac{2+5+3+2+0+4+1+7+4+2+1+0+2}{13} = \mathbf{2.54} \quad \textbf{fallos/mes}$$

$$Mo = \mathbf{2} \quad \textbf{fallos/mes}$$

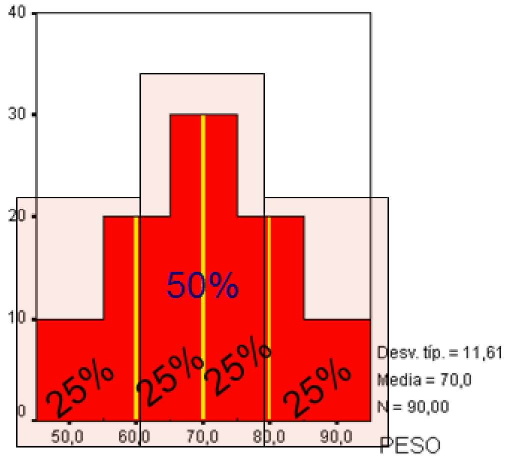
$$Q_1 = x_{\{4\}} = \mathbf{1} \quad \textbf{fallo/mes}$$

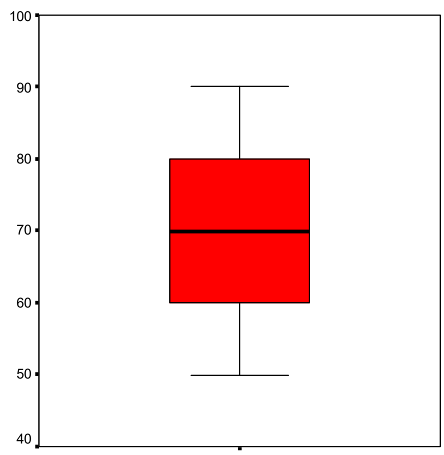
$$Q_3 = x_{\{10\}} = \mathbf{4} \quad \textbf{fallos/mes}$$

Valores	0	1	2	3	4	5	6	7
Frecuencia absoluta	2	2	4	1	2	1	0	1
Frecuencia relativa	0.154	0.154	0.307	0.077	0.154	0.077	0	0.077
Frecuencia relativa acumulada	0.154	0.308	0.615	0.692	0.846	0.923	0.923	1

$$\bar{x} = \sum_{i=1}^{k}[f_i \cdot x_i] = 0.154\cdot 0 + 0.154\cdot 1 + 0.307\cdot 2 + 0.077\cdot 3 + 0.154\cdot 4 + 0.077\cdot 5 + 0.077\cdot 7 = \mathbf{2.54} \quad \textbf{fallos/mes}$$

- Recóllese información sobre o peso de 90 rapaces nunha clase de Matemáticas. Determinar os centís que nos permiten realizar o gráfico de caixa.





– Primeiro cuartil = percentil 25 = 60 Kg.

– Terceiro cuartil= percentil 75= 80 kg.

Actividades propostas

20. Debuxar un diagrama de caixa coñecendo os seguintes datos.

Mínimo valor = 2; cuartil 1 = 3; mediana = 6; cuartil 3 = 7; máximo valor = 12.

21. Un corredor de maratón adestra, de luns a venres percorrendo as seguintes distancias: 2, 3, 3, 6 e 4, respectivamente. Se o sábado tamén adestra:

a) Cantos quilómetros debe percorrer para que a media sexa a mesma?

b) E para que a mediana non varíe?

c) E para que a moda non varíe?

22. O salario mensual en euros dos 6 traballadores dunha empresa téxtil é o que se presenta. Cal dos tres tipos de medidas de tendencia central describe mellor os soldos da empresa?

1 700	1 400	1 700	1 155	1 340	4 565

23. Que valor ou valores poderiamos engadir a este conxunto de valores da variable para que a mediana siga sendo a mesma?

12	19	24	23	23	15	21	32	12	6	32	12	12	21

24. Saen 25 prazas para un posto de auxiliar de enfermaría e preséntanse 200 persoas coas seguintes notas.

notas	3	4	5	6	7	8	9	10
n_i	6	34	25	56	29	10	30	10

a) Con que nota se obtén unha das prazas mediante o exame?

b) Que percentil é a nota 5?

6. MEDIDAS DE DISPERSIÓN

6.1. Medidas de desviacións

As medidas de tendencia central resultan insuficientes á hora de describir unha mostra. Ademais das tendencias, é necesario dispoñer de medidas sobre a variabilidade dos datos. Dentro destas medidas, imos estudar as medidas de desviacións e os rangos.

As medidas de desviacións recollen as desviacións dos valores da variable respecto dunha medida de tendencia central.

A ***varianza*** defínese como:

$$s^2 = \frac{\sum_{i=1}^{N}\left[\left(x_i - \overline{x}\right)^2\right]}{N} = \frac{\sum_{i=1}^{N} x_i^{\,2}}{N} - \overline{x}^2$$

As súas principais vantaxes son a súa manexabilidade matemática e que utiliza todas as observacións. Os seus principais inconvenientes son ser moi sensible a observacións extremas e que a súa unidade é o cadrado da unidade orixinal da mostra.

A **desviación típica** é a raíz cadrada da varianza e ten a principal vantaxe de que utiliza as mesmas unidades que os valores da variable orixinais.

Observa que a desviación típica é unha distancia, a distancia dos valores da variable á media. Recorda que a raíz cadrada é sempre un número positivo.

Asociado á media e á desviación típica, defínese o **coeficiente de variación**, definido en mostras con media distinta de cero como:

$$g = \frac{s}{|\overline{x}|}$$

Este coeficiente é adimensional (non ten unidades e sóese expresar en porcentaxe), o que resulta unha gran vantaxe, xa que permite comparar a variabilidade de distintas mostras, independentemente das súas unidades de medida. Algúns autores definen este coeficiente utilizando a media no denominador, en lugar do seu valor absoluto. Valores do coeficiente de variación maiores do 100% indican que a media non se pode considerar representativa do conxunto de valores da variable.

Exemplo:

- A nota media de 6 alumnos dunha mesma clase de 4º ESO en Matemáticas é de 5. Se a varianza é 0.4, a desviación típica será de 0.632, polo tanto, a media é bastante homoxénea na distribución. As notas que se obtiveron están situadas arredor da nota media 5.

Actividades resoltas

- O propietario dunha instalación mixta solar-eólica está realizando un estudo do volume de enerxía que é capaz de producir a instalación. Para iso, mide esta enerxía ao longo dun total de *N*=16 días que considera suficientemente representativos. A enerxía (en quilovatio, KWh) producida nestes días por dúas instalacións encóntrase recollida na seguinte táboa:

Xeración solar	13.1	10.5	4.1	14.8	19.5	11.9	18	8.6
Xeración eólica	8.5	14.3	24.7	4	2.3	6.4	3.6	9.2
Xeración solar	5.7	15.9	11.2	6.8	14.2	8.2	2.6	9.7
Xeración eólica	13.5	1.4	7.6	12.8	10.3	16.5	21.4	10.9

Utilizando estes valores da variable calcula as medidas de dispersión estudadas, comparando os resultados nas dúas instalacións

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{13.1+10.5+4.1+14.8+19.5+11.9+18+8.6+5.7+15.9+11.2+6.8+14.2+8.2+2.6+9.7}{16} = 10.925 \text{ Kwh}$$

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{8.5+14.3+24.7+4+2.3+6.4+3.6+9.2+13.5+1.4+7.6+12.8+10.3+16.5+21.4+10.9}{16} = 10.463 \text{ Kwh}$$

$$s_x{}^2 = \frac{\sum_{i=1}^{N} x_i{}^2}{N} - \bar{x}^2 = \frac{13.1^2+10.5^2+4.1^2+14.8^2+19.5^2+11.9^2+18^2+8.6^2+5.7^2+15.9^2+11.2^2+6.8^2+14.2^2+8.2^2+2.6^2+9.7^2}{16}$$

$$= \frac{141.5}{16} - 10.9^2 = 22.16$$

$$s_y{}^2 = \frac{\sum_{i=1}^{N} y_i{}^2}{N} - \bar{y}^2 = \frac{8.5^2+14.3^2+24.7^2+4^2+2.3^2+6.4^2+3.6^2+9.2^2+13.5^2+1.4^2+7.6^2+12.8^2+10.3^2+16.5^2+21.4^2+10.9^2}{16}$$

$$- 10.5^2 = \frac{150.48}{16} - 10.5^2 = 41.01$$

$$g_x = \frac{s_x}{|\bar{x}|} = \frac{\sqrt{22.16}}{10.9} = \frac{4.7}{10.9} = 0.43$$

$$g_y = \frac{s_y}{|\bar{y}|} = \frac{\sqrt{41.01}}{10.5} = \frac{6.4}{10.5} = 0.61$$

A media da primeira instalación é máis representativa que a media da segunda xa que o coeficiente de variación é menor na primeira. Os datos están menos agrupados na segunda das instalacións. A súa desviación típica é moito maior.

- Estase realizando un control de calidade sobre os fallos dunhas determinadas máquinas. Para iso, contabilízanse os fallos de *N* = 13 máquinas durante un mes, obtendo os seguintes números de fallos. Utilizando estes valores presentados na táboa de frecuencias obter as medidas de dispersión estudadas.

Valores	0	1	2	3	4	5	6	7
Frecuencia absoluta	2	2	4	1	2	1	0	1
Frecuencia relativa	0.154	0.154	0.307	0.077	0.154	0.077	0	0.077
Frecuencia relativa acumulada	0.154	0.308	0.615	0.692	0.846	0.923	0.923	1

$$\overline{x} = \frac{\sum_{i=1}^{N} [x_i]}{N} = \frac{2+5+3+2+0+4+1+7+4+2+1+0+2}{13} = \mathbf{2.54} \quad \textbf{fallos/mes}$$

$$S^2 = \sum_{i=1}^{k} \left[f_i \cdot (x_i - \overline{x})^2 \right] = 0.154 \cdot (-2.54)^2 + 0.154 \cdot (-1.54)^2 + 0.307 \cdot (-0.54)^2 + 0.077 \cdot 0.46^2 + 0.154 \cdot 1.46^2 +$$

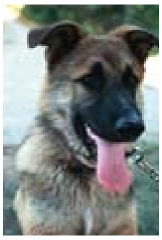
$$+ 0.077 \cdot 2.46^2 + 0.077 \cdot 4.46^2 = \mathbf{3.80} \quad (\textbf{fallos/mes})^2$$

Outra forma de realizar estes mesmos cálculos é:

Valores	0	1	2	3	4	5	6	7	Suma
Frecuencia absoluta	2	2	4	1	2	1	0	1	13
x_i^2	0	1	4	9	16	25	36	49	
x_i^2 · Fr. Abs.	0	2	16	9	32	25	0	49	133

Aplicamos a fórmula: $s^2 = \frac{\sum_{i=1}^{N} x_i{}^2}{N} - \overline{x}^2$ e obtemos que

$s^2 = 133/13 - 2.54^2 = 10.23 - 6.45 = 3.80$, polo que $s = 1.95$.



Actividades propostas

25. Un grupo de cans pastor alemán ten unha media de 70 kg e desviación típica 2 kg. Un conxunto de cans caniche ten unha media de 15 kg e desviación típica 2 kg. Compara ambos os grupos.

26. O tempo, en minutos, que un conxunto de estudantes de 4º ESO dedica a preparar un exame de Matemáticas é:

234	345	345	123	234	234	556
234	234	345	223	167	199	490

As cualificacións dese conxunto de estudantes son as seguintes:

4	5	6	7	6	5	8
9	8	7	8	7	6	8

a) Que teremos que facer para comparar a súa variabilidade? b) En que conxunto os valores da variable están máis dispersos? c) É a media sempre maior que a desviación típica?

6.2. Os rangos

Estas medidas proporcionan información sobre o intervalo total de valores que toma a mostra analizada.

O **rango total** ou **percorrido** é a diferenza entre os valores máximos e mínimos que toma a variable na mostra:

$$R = x_{\{N\}} - x_{\{1\}}$$

O **percorrido intercuartílico** é a diferenza entre o terceiro e o primeiro cuartil:

$$R_I = Q_3 - Q_1$$

Exemplo:

- Estase realizando un control de calidade sobre os fallos dunha determinada máquina. Para iso, contabilízanse os fallos de N=13 máquinas durante un mes, obtendo os seguintes números de fallos: 2, 5, 3, 2, 0, 4, 1, 7, 4, 2, 1, 0, 2. Utilizando estes valores obtemos o rango total igual a 7 e o percorrido intercuartílico igual a 3.

Actividades resoltas

- Saen 25 prazas para un posto de caixeiro nun supermercado e preséntanse 200 persoas. A seguinte información recolle as notas dun test de coñecementos básicos.

notas	2	3	4	5	6	7	8	9	10
n_i	6	4	30	25	56	29	10	30	10

Calcula o rango total da variable obxecto de estudo.

Actividades propostas

27. Recolleuse unha mostra de 20 recipientes cuxos diámetros son:

0.91	1.04	1.01	1	0.77	0.78	1	1.3	1.02	1
1	0.88	1.26	0.92	0.98	0.78	0.82	1.2	1.16	1.14

a) Calcula todas as medidas de dispersión que coñezas.

b) A partir de que valor de diámetro dos recipientes se consideran o 20% con maior diámetro?

7. DISTRIBUCIÓNS BIDIMENSIONAIS

Este apartado céntrase na análise de datos bidimensional, no que son dúas as variables de interese. Deste modo, cando se está analizando unha poboación e se selecciona unha mostra, para cada individuo tómanse dous valores, correspondentes a dúas características (ou variables) distintas. Neste sentido, pode ser interesante considerar simultaneamente os dous caracteres a fin de estudar as posibles relacións entre eles.

7.1. Táboas de frecuencia dunha variable bidimensional

Cando se queren resumir os resultados dunha mostra bidimensional utilizando unha táboa de frecuencias (xa sexa por tratarse dunha variable discreta ou porque se desexen agrupar as observacións dunha variable continua), é preciso utilizar o que se denomina *táboa de dobre entrada* (ou bidimensional). Sexan $x_1, x_2, \ldots, x_k$ as modalidades da primeira variable e $y_1, y_2, \ldots, y_p$ as da segunda. Estas modalidades poden corresponder tanto aos valores que se dan na mostra (se a variable é discreta), como ás marcas de clase dos intervalos utilizados (se a variable é continua). Para construír a táboa de frecuencias, utilízanse as frecuencias absolutas n_{ij} correspondentes ás observacións que toman simultaneamente valores correspondentes ás clases x_i e y_j. Obviamente, hase verificar que:

$$\sum_{i=1}^{k}\sum_{j=1}^{p}\left[n_{ij}\right] = N$$

Con isto, a táboa de frecuencias absolutas preséntase como:

	y_1	y_2	……	y_p	$n_{i\cdot}$
x_1	n_{11}	n_{12}	……	n_{1p}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	……	n_{2p}	$n_{2\cdot}$
……	……	……	……	……	……
x_k	n_{k1}	n_{k2}	……	n_{kp}	$n_{k\cdot}$
$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	……	$n_{\cdot p}$	N

Os valores $n_{i\cdot}$ recollen as frecuencias absolutas da clase x_i, mentres que n_j é a suma de frecuencias absolutas da clase y_j., co que se verifica:

$$n_{i\cdot} = \sum_{j=1}^{p}\left[n_{ij}\right] \qquad n_{\cdot j} = \sum_{i=1}^{k}\left[n_{ij}\right]$$

$$\sum_{i=1}^{k}\left[n_{i\cdot}\right] = N \qquad \sum_{j=1}^{p}\left[n_{\cdot j}\right] = N$$

Da mesma maneira, pódese realizar unha táboa de frecuencias relativas f_{ij}, utilizando os cocientes entre as frecuencias absolutas e o número de observacións:

$$f_{ij} = \frac{n_{ij}}{N} \leq 1$$

Actividades resoltas

- O propietario dunha instalación mixta solar-eólica está realizando un estudo do volume de enerxía que é capaz de producir a instalación. Para iso, mide esta enerxía ao longo dun total de *N=16* días que considera suficientemente representativos. A enerxía (en *kWh*) producida neses días polas instalacións solar e eólica pódese resumir nas seguintes táboas de dobre entrada de frecuencias absolutas e de frecuencias relativas:

		Enerxía eólica				
		[0,6.5]	(6.5,13]	(13,19.5]	(19.5,26]	$n_{i\cdot}$
Enerxía solar	[0,5]	0	0	0	2	**2**
	(5,10]	0	3	2	0	**5**
	(10,15]	2	3	1	0	**6**
	(15,20]	3	0	0	0	**3**
	$n_{\cdot j}$	**5**	**6**	**3**	**2**	**16**

		Enerxía eólica				
		[0,6.5]	(6.5,13]	(13,19.5]	(19.5,26]	$f_{i\cdot}$
Enerxía solar	[0,5]	0	0	0	0.125	**0.125**
	(5,10]	0	0.1875	0.125	0	**0.3125**
	(10,15]	0.125	0.1875	0.0625	0	**0.375**
	(15,20]	0.1875	0	0	0	**0.1875**
	$f_{\cdot j}$	**0.3125**	**0.375**	**0.1875**	**0.125**	**1**

7.2. Representación gráfica dunha variable bidimensional

Ao igual que no caso dunha mostra unidimensional, en numerosas ocasións resulta interesante realizar unha representación gráfica dunha mostra bidimensional.

Un modo sinxelo de representar unha mostra bidimensional é mediante o denominado **diagrama de dispersión** ou **nube de puntos**. Esta técnica consiste en representar no plano (*x,y*) os valores obtidos na mostra.

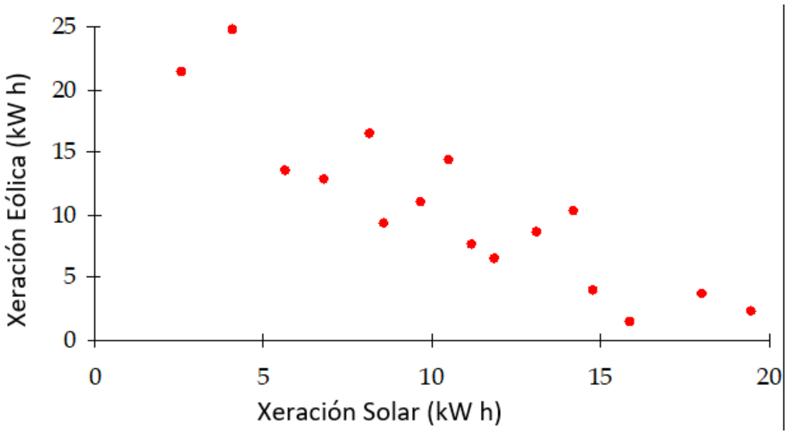


Diagrama de dispersión da xeración solar e eólica (en kWh) da actividade resolta

A figura anterior amosa o diagrama de dispersión. Pódese observar a existencia dunha dependencia inversa.

7.3. Medidas nunha variable bidimensional. Coeficiente de correlación

Cando se está analizando unha mostra bidimensional, pódense calcular as medidas que caracterizan a cada unha das variables da mostra por separado, tal e como se describiu anteriormente. Pero neste caso pódese dar un paso máis e calcular algunhas medidas conxuntas que teñen en conta simultaneamente os valores que toman ambas as variables en cada individuo.

Ao igual que cando se analiza unha única característica, suporemos que se toma unha mostra de tamaño *N* da poboación, é dicir, que está composta por *N* individuos (ou observacións), dos cales se desexa analizar as características (ou variables) *X* e *Y*. Isto dá lugar á obtención de *N* valores para cada unha das dúas variables: (x_1, y_1), (x_2, y_2),…, (x_N, y_N). De novo, estes valores non se supoñen ordenados, senón que o subíndice indica a orde na que foron seleccionados.

Seguindo esta notación pódense formular os cálculos dos momentos respecto á orixe e respecto á media para unha variable bidimensional. Definimos, polo tanto:

Momentos respecto á orixe da orde (*r*,*s*) como:

$$a_{r,s} = \frac{\sum_{i=1}^{N}\left[x_i^{\,r} \cdot y_i^{\,s}\right]}{N}$$

Observa que os momentos respecto á orixe da orde (1, 0) e (0, 1) coinciden coas medias de ambas as variables:

$$a_{1,0} = \overline{x} \qquad a_{0,1} = \overline{y}$$

Tamén resulta de interese ao momento da orde (1,1):

$$a_{1,1} = \frac{\sum_{i=1}^{N}\left[x_i \cdot y_i\right]}{N}$$

Analogamente, pódense definir os momentos respecto á media da orde (*r*,*s*):

$$m_{r,s} = \frac{\sum_{i=1}^{N}\left[\left(x_i - \overline{x}\right)^r \cdot \left(y_i - \overline{y}\right)^s\right]}{N}$$

Os momentos respecto á media da orde (2, 0) e (0, 2) coinciden coas varianzas de ambas as variables:

$$m_{2,0} = s_X^{\,2} \qquad m_{0,2} = s_Y^{\,2}$$

O momento respecto á media da orde (1,1), que se denomina **covarianza** ou momento mixto, é de gran importancia:

$$m_{1,1} = \frac{\sum_{i=1}^{N}\left[\left(x_i - \overline{x}\right)\cdot\left(y_i - \overline{y}\right)\right]}{N}$$

Alternativamente á fórmula anterior, a covarianza pódese calcular a partir dos momentos respecto á orixe, segundo a fórmula:

$$m_{1,1} = a_{1,1} - a_{1,0} \cdot a_{0,1} = \frac{\sum_{i=1}^{N}(x_i \cdot y_i)}{N} - \bar{x} \cdot \bar{y}$$

A covarianza, ao igual que a varianza, ten o inconveniente de que depende das unidades da mostra.

Por este motivo, utilízase o coeficiente de **correlación** lineal de Pearson (que se denota, indistintamente, como ρ ou r):

$$\rho = r = \frac{m_{1,1}}{s_x \cdot s_y} = \frac{\dfrac{\sum_{i=1}^{N}(x_i \cdot y_i)}{N} - \bar{x} \cdot \bar{y}}{s_x \cdot s_y}$$

Este coeficiente terá o signo da covarianza e indicaranos se a dependencia entre as dúas variables obxecto de estudo son dependentes positiva ou negativamente. O coeficiente de correlación (ou simplemente correlación) toma un valor comprendido entre –1 e 1. Se a correlación é positiva dise que existe dependencia directa entre *X* e Y (a un aumento dunha das dúas variables correspóndelle unha tendencia ao aumento na outra). En cambio, se a correlación é negativa, dise que existe unha dependencia inversa (a un aumento dunha das dúas variables correspóndelle unha tendencia ao decremento na outra).

Actividades resoltas

- O propietario dunha instalación mixta solar-eólica está realizando un estudo do volume de enerxía que é capaz de producir a instalación. Para iso, mide esta enerxía ao longo dun total de *N*=16 días que considera suficientemente representativos. A enerxía (en kWh) producida neses días polas instalacións solar e eólica encóntrase recollida na seguinte táboa:

Xeración solar (x_i)	13.1	10.5	4.1	14.8	19.5	11.9	18	8.6	5.7	15.9	11.2	6.8	14.2	8.2	2.6	9.7
Xeración eólica (y_i)	8.5	14.3	24.7	4	2.3	6.4	3.6	9.2	13.5	1.4	7.6	12.8	10.3	16.5	21.4	10.9

Utilizando estas producións, imos calcular a covarianza e o coeficiente de correlación, denotando a xeración solar como variable ***X*** e a xeración eólica como variable ***Y***. Engadimos novas filas á nosa táboa:

Xeración solar (x_i)	13.1	10.5	4.1	14.8	19.5	11.9	18	8.6	5.7	15.9	11.2	6.8	14.2	8.2	2.6	9.7
Xeración eólica (y_i)	8.5	14.3	24.7	4	2.3	6.4	3.6	9.2	13.5	1.4	7.6	12.8	10.3	16.5	21.4	10.9
x_i^2	171.6	110.3	16.81	219.0	380.3	141.6	324	73.96	32.49	252.8	125.4	46.24	201.6	67.24	6.76	94.09
y_i^2	72.25	204.5	610.1	16	5.29	40.96	12.96	84.64	182.3	1.96	57.76	163.8	106.1	272.3	457.9	118.8
$x_i * y_i$	111.4	150.2	101.3	59.2	44.85	76.16	64.8	79.12	76.95	22.26	85.12	87.04	146.2	135.3	55.64	105.7

Previamente calculamos a media e a desviación típica de cada variable (que xa coñecemos dunha actividade resolta anterior). Sumando a primeira fila e dividindo por N = 16, obtemos a media da Xeración Solar en Kwh. Recorda $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$; polo tanto

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{13.1+10.5+4.1+14.8+19.5+11.9+18+8.6+5.7+15.9+11.2+6.8+14.2+8.2+2.6+9.7}{16} = 10.925 \text{ Kwh}$$

Sumando a segunda fila e dividindo por N = 16 obtemos a media da Xeración Eólica en Kwh:

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{8.5+14.3+24.7+4+2.3+6.4+3.6+9.2+13.5+1.4+7.6+12.8+10.3+16.5+21.4+10.9}{16} = 10.463 \text{ Kwh}$$

Na terceira fila calculamos os cadrados dos valores da primeira variable e utilizámolos para calcular a varianza: Recorda $s_x{}^2 = \frac{\sum_{i=1}^{N} x_i{}^2}{N} - \bar{x}^2$; polo tanto

$$s_x{}^2 = \frac{\sum_{i=1}^{N} x_i{}^2}{N} - \bar{x}^2 = \frac{13.1^2 + 10.5^2 + 4.1^2 + 14.8^2 + 19.5^2 + 11.9^2 + 18^2 + 8.6^2 + 5.7^2 + 15.9^2 + 11.2^2 + 6.8^2 + 14.2^2 + 8.2^2 + 2.6^2 + 9.7^2}{16} = \frac{141.5}{16} - 10.9^2 = 22.16$$

Na cuarta fila calculamos os cadrados dos valores da segunda variable e calculamos a súa varianza tal que

$$s_y{}^2 = \frac{\sum_{i=1}^{N} y_i{}^2}{N} - \bar{y}^2 = \frac{8'5^2 + 14'3^2 + 24'7^2 + 4^2 + 2'3^2 + 6'4^2 + 3'6^2 + 9'2^2 + 13'5^2 + 1'4^2 + 7'6^2 + 12'8^2 + 10'3^2 + 16'5^2 + 21'4^2 + 10'9^2}{16} = \frac{150'48}{16} - 10'5^2 = 41'01$$

A desviación típica é a raíz cadrada da varianza, polo tanto:

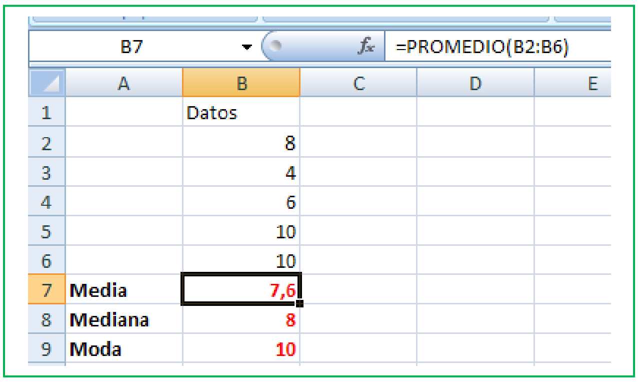
$$s_x = \sqrt{22.16} = 4.71 \text{ e } s_y = \sqrt{41.01} = 6.4$$

Para calcular o coeficiente de correlación calculamos na quinta fila os produtos da variable *x* pola variable *y*. Así, 13.1*8.5 = 111.4. Queremos calcular o termo: $\frac{\sum_{i=1}^{N} (x_i \cdot y_i)}{N}$. Ao sumar obtemos 1 401.2, que dividimos entre 16, restámoslle o produto das medias e dividimos polo produto das desviacións típicas:

$$\rho = \frac{\frac{\sum_{i=1}^{N}(x_i \cdot y_i)}{N} - \bar{x} \cdot \bar{y}}{s_x \cdot s_y} = \frac{1\,401.2 - (10.9 \cdot 10.5)}{4.71 \cdot 6.4} = \frac{-26.728}{4.71 \cdot 6.4} = -0.887$$

Este coeficiente de correlación negativo e achegado a –1 indícanos que a relación entre as dúas variables é negativa e bastante importante.

Utiliza o ordenador



B7 | f_x =PROMEDIO(B2:B6)

	A	B	C	D	E
1		Datos			
2		8			
3		4			
4		6			
5		10			
6		10			
7	Media	7,6			
8	Mediana	8			
9	Moda	10			

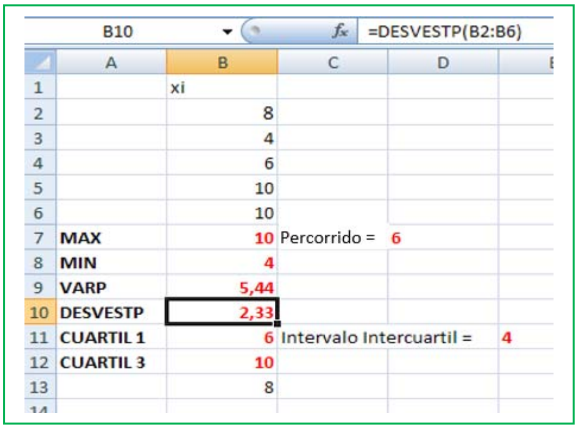
- *Neves tivo en Matemáticas as seguintes notas: 8, 4, 6, 10 e 10. Calcula a súa media, a súa moda e a súa mediana.*

Para calcular a media, a mediana e a moda coa folla de cálculo, copiamos na casa B2, B3… os datos: 8, 4, 6, 10 e 10. Escribimos na casa A7, Media, e para calcular a media escribimos un signo igual en B7. Buscamos, despregando as posibles funcións, a función PROMEDIO, e escribimos

=PROMEDIO(B2:B6),

que significa que calcule a media dos valores que hai nas casas desde B2 ata B6.

Do mesmo modo calculamos a mediana buscando nas funcións ou escribindo =MEDIANA(B2:B6) e a moda buscando nas funcións ou escribindo =MODA(B2,B6).



B10 | f_x =DESVESTP(B2:B6)

	A	B	C	D
1		xi		
2		8		
3		4		
4		6		
5		10		
6		10		
7	MAX	10	Percorrido =	6
8	MIN	4		
9	VARP	5,44		
10	DESVESTP	2,33		
11	CUARTIL 1	6	Intervalo Intercuartil =	4
12	CUARTIL 3	10		
13		8		

Igual que calculamos a media, a mediana e a moda, a folla de cálculo pódese utilizar para obter:

- O percorrido calculando MAX – MIN → 6.
- A varianza utilizando VARP → 5.44.
- A desviación típica usando DESVESTP → 2.33
- Os cuartís, (CUARTIL), sendo o cuartil 0 o mínimo; o cuartil 1, Q1; o cuartil 2, a mediana; o cuartil 3, Q3; e o cuartil 4, o máximo.
- Q1 = 6.
- Q3 = 10.
- O intervalo intercuartílico = 10 – 6 = 4.

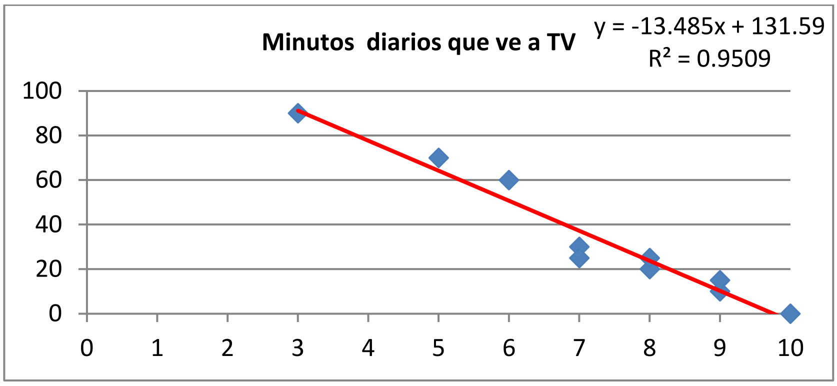
Utiliza o ordenador

- Preguntamos a 10 alumnos de 4º ESO polas súas cualificacións en Matemáticas, polo número de minutos diarios que ven a televisión, polo número de horas semanais que dedican ao estudo, e pola súa estatura en centímetros. Os datos recóllense na táboa adxunta. Queremos debuxar as nubes de puntos que os relacionan coas cualificacións de Matemáticas, o coeficiente de correlación e a recta de regresión.

Cualificacións de Matemáticas	10	3	7	8	5	9	9	8	6	7
Minutos diarios que ve a TV	0	90	30	20	70	10	15	25	60	25
Horas semanais de estudo	15	2	9	12	7	14	13	11	7	8
Estatura (en cm)	177	168	157	159	163	179	180	175	169	170

Para facelo, entramos en Excel, e copiamos os datos. Seleccionamos a primeira e a segunda fila, logo a primeira e a terceira e por último a primeira fila e a cuarta.

Coa primeira e segunda filas seleccionadas, imos *Inserir*, *Dispersión* e eliximos a *nube de puntos*. Podemos conseguir que o eixe de abscisas vaia de 0 a 10 en "*Dar formato ao eixe*". Pinchamos sobre un punto da nube e eliximos "*Agregar liña de tendencia*". Para que debuxe o ordenador a recta de regresión a liña de tendencia debe ser *Lineal*. Na pantalla que aparece marcamos a casa que di: "*Presentar ecuación no gráfico*" e a casa que di "*Presentar o valor de R cadrado no gráfico*".



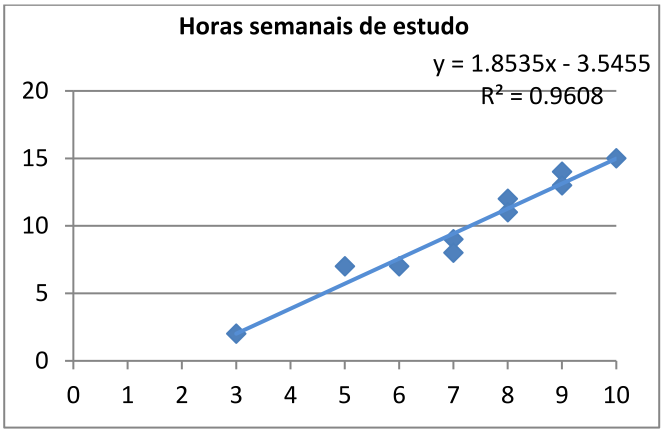
Observa: a recta de regresión, en cor vermella, é decrecente e a súa ecuación é aproximadamente:

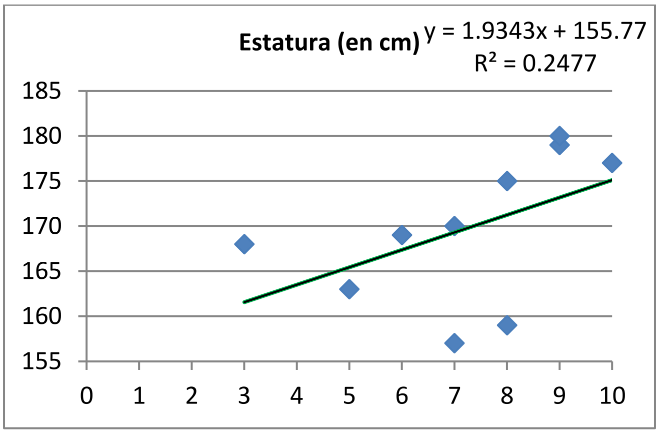
$$y = -13.5\,x + 132.$$

O cadrado do coeficiente de correlación é $\rho^2 = 0.95$. A correlación é negativa e alta:

$$\rho = \sqrt{0.95} = -0.975$$

Facemos o mesmo coa primeira e terceira fila e coa primeira e cuarta fila. Obtemos os gráficos:





Observa que en ambos os casos a pendente da recta de regresión é positiva pero, no primeiro, o coeficiente de correlación, positivo, é próximo a 1, $= \sqrt{0.96} = 0.98$. A correlación é alta e positiva.

No segundo $\rho = \sqrt{0.25} = 0.5$.

Actividades propostas

28. Medíronse os pesos e alturas de 6 persoas, como mostra das persoas que están nunha fila ou cola de espera, obténdose os seguintes resultados:

Pesos (kg)	65	60	65	63	68	68
Alturas (cm)	170	150	168	170	175	180

Pídese:

a) Calcular as medias e as varianzas deses dous conxuntos de datos unidimensionais.

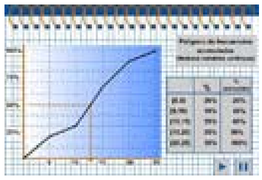
b) Que medidas están máis dispersas, os pesos ou as alturas?

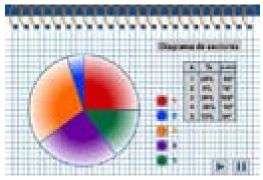
c) Representar graficamente ese conxunto de datos bidimensional. Calcular a covarianza e interpretar o seu valor.

d) Dar unha medida da correlación entre ambas as variables. Interpretar o seu valor.

CURIOSIDADES. REVISTA

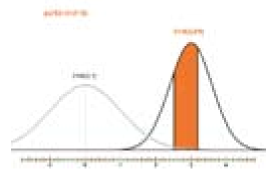
UTILIZAMOS A ESTATÍSTICA POR ENRIBA DAS NOSAS POSIBILIDADES?





Nas últimas décadas o uso de datos estatísticos é unha das principais maneiras coas que se presenta información de calquera tipo, proveña a súa fonte dos medios de comunicación, a través de mensaxes publicitarias ou relacionada con traballos de investigación. Actualmente consumir información convértese, en moitas ocasións, en entrar nun mundo de números, porcentaxes, gráficos, probabilidades, mapas e outros conceptos básicos desta disciplina que custa entender.

"TEÑO OS MEUS RESULTADOS HAI TEMPO, PERO NON SEI COMO CHEGAR A ELES"



Esta expresiva oración de Gauss -descubridor da campá que leva o seu nome, e que alude á distribución normal cando a cantidade de datos é bastante grande-, é aplicable a moitas das informacións erróneas que vemos a diario. Teñen os datos pero non saben como chegar ao núcleo da súa interpretación.

Moitas veces cando un medio de comunicación quere impresionar mediante un titular sobre a gravidade dunha situación que afecta a toda a poboación, fai uso de números absolutos en lugar de porcentaxes.

EL PAÍS

PORTADA | INTERNACIONAL | POLÍT

POLÍTICA

MONCLOA | FERRAZ | GÉNOVA | + PARTIDOS | CONGRESO | OPINIÓN | SUCESOS

ESTÁ PASANDO | El órdago soberanista catalán | La imputación de la Infanta | Baltasar Garzón

OPERACIÓN RETORNO

40 muertos en carretera en Semana Santa

- Los principales atascos de la jornada se concentran en la zona centro
- Consulta el estado de las carreteras

EL PAÍS | Madrid | 9 ABR 2012 - 09:47 CET | 52

Por exemplo: Cando lemos o titular que dúbida cabe que todos pensamos que 40 mortos son moitos mortos sexan por accidente de tráfico ou por outra causa. A argucia está ben pensada para chamar a *atención do* lector pero, informativamente falando, esta presentación dos feitos utilizando números sen comparalos con outros números merece *"un suspenso"*. Os datos estatísticos non "falan por si mesmos". Un dato sempre hai que relacionalo con outros datos para comprender a variabilidade que experimentou o caso que estamos analizando. Se a noticia se tivese acompañado coas estatísticas de mortes por accidente de tráfico dos últimos anos en períodos de vacacións de catro días, rapidamente o lector se daría conta de que non é para alarmarse máis que outras veces xa que o número de mortos nin subiu nin baixou, é máis ou menos o mesmo que en calquera outra ponte similar en días. É dicir, este "abraiante" titular apoiado en datos numéricos, en realidade *nin sequera é noticia*…

RESUMO

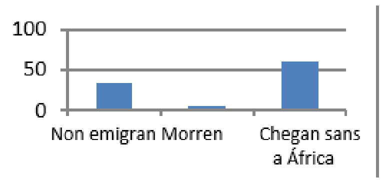
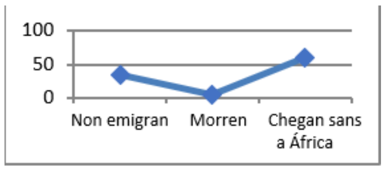
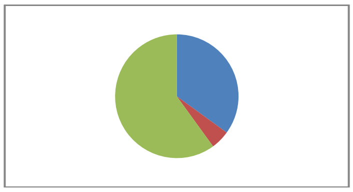
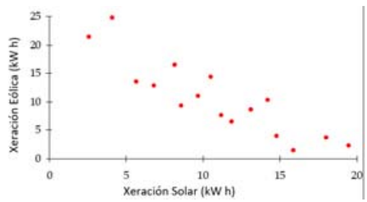
Noción	*Definición*	*Exemplos*
Poboación estatística, colectivo ou universo	O conxunto de todos os individuos (persoas, obxectos, animais, etc.) que conteñan información sobre o fenómeno que se estuda.	Número de persoas en España entre 16-65 anos.
Mostra	É un subconxunto representativo que se selecciona da poboación e sobre o que se vai realizar a análise descritiva. O tamaño da mostra é o número dos seus elementos. Cando a mostra comprende todos os elementos da poboación, denomínase censo.	Número de persoas nun barrio de Madrid entre 16 e 65 anos.
Variable observable ou estatística *X*	En xeral, suporemos que se está analizando unha determinada poboación, da que nos interesa certa característica que vén dada pola variable *X*.	As variables que están baixo estudo pódense clasificar en dúas categorías: Variables cualitativas ou atributos (datos non métricos). Variables cuantitativas, que teñen un valor numérico.
Frecuencia absoluta	Número de veces que se repite un valor da variable	Se ao tirar un dado obtivemos 2 veces o 3, 2 é a frecuencia absoluta de 3.
Frecuencia relativa	Frecuencia absoluta dividido polo número de experimentos	Se se realiza un experimento 500 veces e a frecuencia absoluta dun suceso é 107, a frecuencia relativa é 107/500.
Frecuencia acumulada	Súmanse as frecuencias anteriores	
Diagrama de rectángulos ou barras	Os valores da variable represéntanse mediante rectángulos de igual base e de altura proporcional á frecuencia. Indícanse no eixe horizontal a variable e no vertical as frecuencias.	
Polígono de frecuencias	Únense os puntos medios superiores dun diagrama de barras.	

Diagrama de sectores	Nun círculo debúxanse sectores de ángulos proporcionais ás frecuencias.	
Media aritmética	É o cociente entre a suma de todos os valores da variable e o número total de datos.	Nos datos 3, 5, 5, 7, 8, a media é: (3 + 5 + 5 + 7 + 8)/5 = 28/5 = 5.6.
Mediana	Deixa por debaixo a metade dos valores e por enriba a outra metade.	A mediana é 5.
Moda	O valor que máis se repite.	A moda é 5.
Varianza	Medida de desviación que recolle as desviacións dos valores da variable respecto da media aritmética.	$s^2 = \frac{\sum_{i=1}^{N}\left[(x_i - \bar{x})^2\right]}{N}$
Desviación típica	A desviación típica é a raíz cadrada da varianza.	
Coeficiente de variación	Permite comparar a variabilidade de distintas mostras, independentemente das súas unidades de medida.	$g = \frac{s}{\lvert\bar{x}\rvert}$
Rango total ou percorrido	Diferenza entre os valores máximos e mínimos que toma a variable na mostra.	$R = x_{\{N\}} - x_{\{1\}}$
Percorrido intercuartílico	Diferenza entre o terceiro e o primeiro cuartil.	$R_I = Q_3 - Q_1$
Nube de puntos	Un modo sinxelo de representar unha mostra bidimensional. Esta técnica consiste en representar no plano (*x, y*) os valores obtidos na mostra.	
Coeficiente de correlación	Indica se a dependencia entre dúas variables obxecto de estudo son dependentes positiva ou negativamente.	$\rho = r = \frac{m_{1,1}}{s_x \cdot s_y}$

EXERCICIOS E PROBLEMAS

Poboación e mostra. Variables estatísticas. Táboas de frecuencias

1. Lánzase unha moeda 700 veces e obtense cara 355 veces. Expresa nunha táboa as frecuencias absolutas, relativas e calcula tamén as frecuencias acumuladas absolutas e acumuladas relativas de caras e cruces neste experimento.

2. Lánzase un dado 500 veces e obtéñense os seguintes resultados:

Resultado	1	2	3	4	5	6
Número de veces	70	81	92	85		81

a) Cantas veces saíu o 5?

b) Construír unha táboa coas frecuencias absolutas e as frecuencias absolutas acumuladas.

c) Construír unha táboa coas frecuencias relativas e as frecuencias relativas acumuladas.

3. Unha urna contén 10 bólas numeradas do 0 ao 9. Sacamos unha bóla, anotamos o número e devolvemos a bóla á urna. Repetimos o experimento 1000 veces e obtivéronse os resultados indicados na táboa:

Resultado	0	1	2	3	4	5	6	7	8	9
Frecuencia absoluta	79	102			93	98	104	77		
Frecuencia relativa			0.12	0.13					0.1	
Frecuencia absoluta acumulada	79	181								
Frecuencia relativa acumulada										1

a) Cal é a frecuencia absoluta de 9?

b) Cal é a frecuencia absoluta acumulada de 2?

c) Cal é a frecuencia relativa acumulada de 1?

d) Copia a táboa no teu caderno e complétaa.

4. Pepa tirou un dado 25 veces e obtivo os seguintes resultados:

1, 2, 5, 6, 3, 1, 4, 5, 6, 1, 3, 1, 2, 2, 1, 6, 2, 2, 4, 3, 4, 6, 6, 1, 4

a) Construír unha táboa de frecuencias absolutas.

b) Construír unha táboa de frecuencias relativas.

c) Debuxa un diagrama de barras.

d) Debuxa un polígono de frecuencias e unha representación por sectores.

5. Nunha clase mediuse o tamaño das mans de cada un dos alumnos e alumnas, e o resultado en centímetros foi o seguinte:

19, 18, 20, 19, 18, 21, 19, 17, 16, 20,

16, 19, 20, 21, 18, 17, 20, 19, 22, 21,

23, 21, 17, 18, 17, 19, 21, 20, 16, 19

a) Que tamaño foi o valor mínimo? E o máximo? Cal é o rango total da variable?

b) Construír unha táboa de frecuencias absolutas e outra de frecuencias relativas.

c) Construír unha táboa de frecuencias absolutas acumuladas e outra de frecuencias relativas acumuladas.

6. Calcula a frecuencia absoluta dos datos dunha enquisa na que se elixiu entre ver a televisión, t, ou ler un libro, l:

t, l, t, t, t, l, t, t, l, t, l, t, l, t, t, t, l, l, t, l, t, l, t, l, t.

7. A duración en minutos dunhas chamadas telefónicas foi:

7, 3, 6, 3, 7, 5, 4, 3, 5, 7, 10, 1, 9, 12, 2

Construír unha táboa de frecuencias absolutas e unha táboa de frecuencias relativas.

Gráficos estatísticos

8. Preguntouse nunha vila da provincia de Madrid o número de irmáns que teñen e obtívose a seguinte táboa de frecuencias absolutas sobre o número de fillos de cada familia:

Número de fillos	1	2	3	4	5	6	7	8 ou máis
Número de familias	46	249	205	106	46	21	15	6

a) Escribe no teu caderno unha táboa de frecuencias relativas.

b) Fai un diagrama de barras de frecuencias absolutas e outro de frecuencias relativas.

c) Fai un polígono de frecuencias absolutas e outro de frecuencias absolutas acumuladas.

9. Fai unha enquisa similar cos teus compañeiros e compañeiras de curso preguntando o número de irmáns e confeccionando unha táboa sobre o número de fillos e o número de familias.

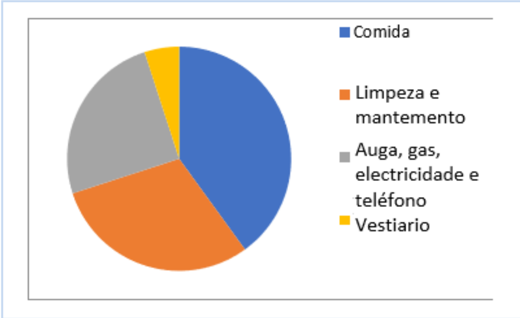
a) Constrúe unha táboa de frecuencias relativas.

b) Fai un diagrama de barras de frecuencias absolutas e relativas. Completa cun polígono de frecuencias.

c) Compara a táboa de frecuencias relativas e o diagrama de barras de frecuencias relativas que obteñas co obtido no exercicio anterior.

10. Un batido de froita contén 25 % de laranxa, 15 % de plátano; 50 % de mazá e o resto de leite. Representa nun diagrama de sectores a composición do batido.

11. Nun campamento de verán gastáronse dez mil euros. O gráfico amosa a distribución do gasto:



1. Comida: 40 %
2. Limpeza e mantemento: 30 %
3. Auga, gas, electricidade e teléfono: 25 %
4. Vestiario: ....

a) Que porcentaxe se gastou en vestiario?

b) Cantos euros se gastaron en comida?

c) Canto mide o ángulo do sector correspondente a actividades?

12. Busca en revistas ou periódicos dúas gráficas estatísticas, recórtaas e pégaas no teu caderno. En moitas ocasións estas gráficas teñen erros. Obsérvaas detidamente e comenta as seguintes cuestións:

a) Está clara a variable á que se refire? E as frecuencias?

b) Son correctas as unidades? Poden mellorarse?

c) Comenta as gráficas.

13. Faise unha enquisa sobre o número de veces que van ao cine uns mozos ao mes. Os valores da variable están na táboa:

Veces que van ao cine	0	1	2	3	4	5
Frecuencia absoluta	1	7	9	5	2	1

a) Representa un diagrama de barras de frecuencias absolutas.

b) Representa un polígono de frecuencias relativas.

c) Representa os valores da variable nun diagrama de sectores.

14. Faise un estudo sobre o que se recicla nunha cidade e faise unha táboa co peso en porcentaxe dos distintos tipos de residuos:

Tipo de residuo	Porcentaxe
Orgánico	15
Papel e cartón	1
Vidro	15
Plástico	1
Pilas	15

a) Constrúe un diagrama de barras

b) Representa un polígono de frecuencias.

c) Representa os valores da variable nun diagrama de sectores.

15. Nun exercicio anterior obtívose o resultado de medir nunha clase o tamaño das mans de cada un dos alumnos e alumnas, e o resultado en centímetros foi o seguinte:

19, 18, 20, 19, 18, 21, 19, 17, 16, 20,

16, 19, 20, 21, 18, 17, 20, 19, 22, 21,

23, 21, 17, 18, 17, 19, 21, 20, 16, 19

Representa os valores da variable nun diagrama de barras e nun polígono de frecuencias.

16. O 35 % das cegoñas non emigrou este ano a África e o 6 % morreu polo camiño. Debuxa un diagrama por sector que describa esta situación.

17. Nunha clase preguntouse polas preferencias deportivas e obtívose:

Fútbol	Baloncesto	Natación	Karate	Ciclismo
8	9	7	6	10

a) Copia a táboa no teu caderno e constrúe unha táboa de frecuencias relativas.

b) Representa estes valores da variable nun diagrama de sectores.

Medidas de centralización e dispersión

18. Pepa tirou un dado 25 veces nun exercicio anterior e obtivo os seguintes resultados:

1, 2, 5, 6, 3, 1, 4, 5, 6, 1, 3, 1, 2, 2, 1, 6, 2, 2, 4, 3, 4, 6, 6, 1, 4

a) Calcula a media aritmética.

b) Calcula a mediana.

c) Cal é a moda? É única?

d) Calcula a varianza e desviación típica interpretando o seu resultado.

19. Sara tivo as seguintes notas nos seus exames de Matemáticas: 9, 7, 8, 6, 9, 10, 9

a) Calcula a media aritmética.

b) Calcula a mediana.

c) Cal é a moda? É única?

d) Calcula o percentil 45 interpretando o seu resultado.

e) Calcula o percentil 75 interpretando o seu resultado. Que outro nome recibe?

f) Calcula a varianza e a desviación típica interpretando o seu resultado.

g) Calcula o coeficiente de variación interpretando o seu resultado.

20. Nun exercicio anterior obtívose o resultado de medir nunha clase o tamaño das mans de cada un dos alumnos e alumnas, e o resultado en centímetros foi o seguinte:

19, 18, 20, 19, 18, 21, 19, 17, 16, 20,

16, 19, 20, 21, 18, 17, 20, 19, 22, 21,

23, 21, 17, 18, 17, 19, 21, 20, 16, 19

a) Calcula a media aritmética.

b) Calcula a mediana.

c) Cal é a moda? É única?

d) Calcula o percentil 45 interpretando o seu resultado.

e) Calcula o percentil 75 interpretando o seu resultado. Que outro nome recibe?

f) Calcula a varianza e desviación típica interpretando o seu resultado.

g) Calcula o coeficiente de variación interpretando o seu resultado.

21. Interésanos coñecer a distribución de notas obtidas por 40 estudantes. As notas son:

4, 1, 7, 10, 3, 2, 8, 9, 0, 0, 5, 8, 2, 7, 1, 2, 8, 10, 2, 10,

3, 4, 8, 9, 3, 6, 3, 7, 2, 4, 9, 4, 9, 5, 1, 3, 3, 9, 7, 8, 10

a) Escribe no teu caderno unha táboa de frecuencias absolutas.

b) Fai un polígono de frecuencias absolutas.

c) Calcula a media.

d) Calcula a mediana.

e) Calcula a moda.

f) Calcula o percentil 45 interpretando o seu resultado.

g) Calcula o percentil 75 interpretando o seu resultado. Que outro nome recibe?

h) Calcula a varianza e desviación típica interpretando o seu resultado.

i) Calcula o coeficiente de variación interpretando o seu resultado.

j) Se as notas dos mesmos alumnos respecto a outra materia teñen unha media de 5.3 e desviación típica de 2, cal das dúas materias ten unha media máis homoxénea?

22. Os xogadores dun equipo de balonmán teñen as seguintes idades:

12, 14, 13, 12, 15, 11, 12, 12, 13, 14, 11, 12, 12.

a. Calcula a media.

b. Calcula a mediana.

c. Calcula a moda.

d. Calcula o percentil 45 interpretando o seu resultado.

e. Calcula o percentil 75 interpretando o seu resultado. Que outro nome recibe?

f. Calcula a varianza e desviación típica interpretando o seu resultado.

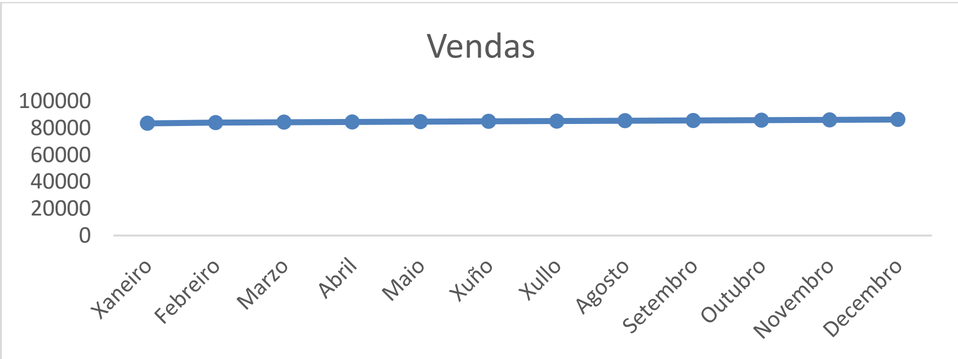
g. Calcula o coeficiente de variación interpretando o seu resultado.

Problemas

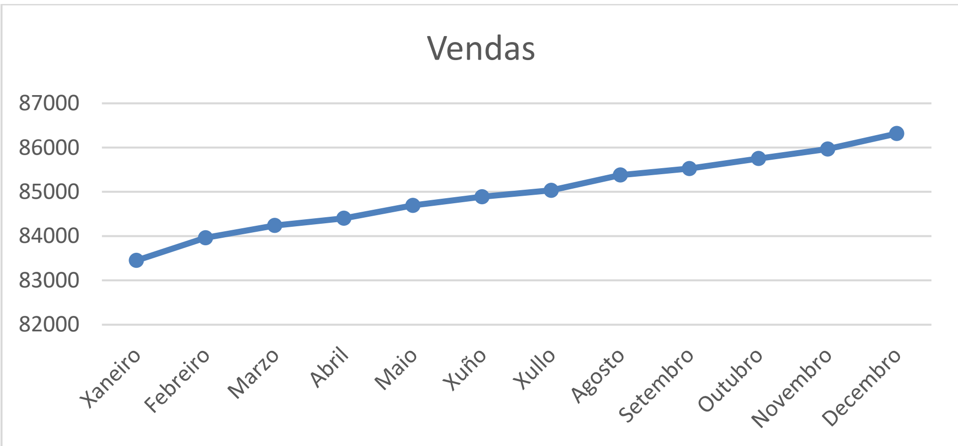
23. O Director Comercial dunha empresa vai ser avaliado. Para iso debe dar conta dos resultados obtidos. Quere quedar ben, pois iso pódelle supoñer un aumento de soldo. Vendeu as seguintes cantidades:

Meses	Xaneiro	Febreiro	Marzo	Abril	Maio	Xuño	Xullo	Agosto	Setembro	Outubro	Novembro	Decembro
Vendas	83451	83962	84238	84401	84693	84889	85032	85378	85524	85751	85967	86316

O estatístico da empresa entregoulle a seguinte gráfica:



Non lle gustou nada e, para a presentación, confecciona el mesmo o seguinte gráfico:



Ambos os gráficos son correctos. Escribe un informe sobre como poden os distintos gráficos dar impresións tan diferentes.

24. Tira unha moeda 15 veces e anota as veces que cae cara e as que non. Constrúe logo dúas táboas: unha de frecuencias absolutas e outra de frecuencias relativas. Representa o resultado nun diagrama de frecuencias e nun polígono de frecuencias.

25. A media de seis números é 5. Engádense dous números máis pero a media segue sendo 5. Canto suman estes dous números?

26. A seguinte táboa expresa as estaturas, en metros, de 1000 soldados:

Talle	1.50 – 1.56	1.56 – 1.62	1.62 – 1.68	1.68 - 1.74	1.74 - 1.80	1.80-1.92
Nº de soldados	20	150	200	330	200	100

Calcula:

a) A media e a desviación típica.

b) Os intervalos onde se encontran a mediana e os cuartís.

c) O intervalo ($\overline{x}$ - σ, $\overline{x}$ + σ) e a porcentaxe de individuos neste intervalo.

d) Representa os datos nun histograma.

27. Unha compañía aérea sospeita que existe unha relación entre as variables *X*, tempo dun voo, en horas; e *Y*, consumo de combustible (gasóleo) para este voo, en litros. Por esta razón, obtivéronse os seguintes datos, dentro do rango de niveis de interese para X nesta compañía.

X_{i}	0.4	0.5	0.6	0.65	0.7	0.8	1	1.15	1.2
Y_{i}	1 350	2 220	2 900	3 150	3 350	3 550	3 900	4 330	4 500

X_i	1.4	1.5	1.6	1.8	2.2	3
Y_i	5 050	5 320	5 650	6 400	7 500	10 250

Pídese:

a) Mediante a representación do diagrama de dispersión razoar o interese de relacionar estas variables.

b) Obter a covarianza e o coeficiente de correlación entre ambas as variables. Interpretar os resultados.

AUTOAVALIACIÓN

1. Un diagrama de caixa informa sobre:

a) Os cuartís e curtosis. b) Asimetría e varianza. c) Datos atípicos e simetría.

2. Sexa a variable aleatoria o número de persoas que é capaz de levantar un ascensor. Para calcular o nº de persoas a partir do cal se recolle o 30% dos valores da variable necesitamos obter

a) O percentil 30 b) o percentil 3 c) o percentil 70

3.O 25% dos madrileños gastan na factura do móbil por enriba de 100 euros mentres que o 25% gastan por debaixo de 20 euros. Entón coñecemos:

a) 100 e 20 son valores que corresponden ao cuartil 1 e 3, respectivamente.

b) 100 e 20 son valores que corresponden ao cuartil 3 e 1, respectivamente.

c) 100 e 20 son valores que non corresponden a ningún cuartil.

4.Nun diagrama de barras de frecuencias absolutas, a suma das súas alturas é proporcional a:

a) 100 b) 1 c) Total de valores da variable d) Suma das súas bases

5. A media dos seguintes valores da variable 3, 4, 6, 7, 5, 8, é:

a) 6 b) 7 c) 4.8 d) 5.5

6.A mediana dos seguintes valores da variable 3, 4, 6, 7, 8, é:

a) 6 b) 7 c) 4 d) 5

7.A moda dos seguintes valores da variable 3, 4, 6, 7, 5, 8, 7, 7, é:

a) 6 b) 7 c) 4 d) 5

8.A media de 7 números é 8. Engádense dous números máis pero a media segue sendo 8. Canto suman estes dous números?

a) 10 b) 16 c) 20 d)14

9. Dúas revistas especializadas en emprego, A e B, publicaron unha media de ofertas de traballo de m_A = 10 e m_B = 20 con varianzas, respectivamente, de $s^2_A = 4$ e $s^2_B = 9$.

a) A revista B presenta maior coeficiente de variación que a revista A.

b) A revista A presenta maior coeficiente de variación que a revista B.

c) A revista B presenta igual coeficiente de variación que A

10. O 70 % dos madrileños gastan en regalos de Nadal por enriba de 100 euros mentres que o 5 % gastan por enriba de 500 euros. Entón coñecemos:

a) O valor correspondente ao percentil 30.

b) O valor correspondente ao percentil 70.

c) O valor correspondente ao percentil 5.